The Dissertation Committee for Jina Kang Certifies that this is the approved version of the following dissertation:

Examining Scientific Thinking Processes in Open-Ended Serious Games through Gameplay Data

Committee:

Min Liu, Supervisor

Paul Resta

Catherine Riegle-Crumb

Lucas Horton

Examining Scientific Thinking Processes in Open-Ended Serious Games through Gameplay Data

by

Jina Kang, B.S.; M.S.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2017

Acknowledgements

I would like to express my deep appreciation to my advisor, Dr. Min Liu, for the scholarly guidance and mentorship she provided to me through the years at the University of Texas at Austin. Dr. Liu has provided me with numerous opportunities to train as a teacher and a mentor, and practice as a researcher. I am also thankful for the opportunity to learn from my committee members, Dr. Paul Resta, Dr. Catherine Riegle-Crumb, and Dr. Lucas Horton. Their perspectives and insights were immensely valuable to me. I have greatly enjoyed collaborating with Dr. Horton over the years in a project team, *Alien Rescue*, and am very thankful that he agreed to join my committee. Thanks as well to Dr. Siew Ang. I am very thankful that she provided me a practical guidance on my studies and encouragement.

I would like to appreciate all graduate student colleagues I have worked with in the team *Alien Rescue*. Special thanks to the members, Sa Liu and Chenglu Li for developing a new version of *Alien Rescue* within a limited time frame and Wenting Zou, Zilong Pan, and Hyeyeon Lee assisting in data analysis. Thanks as well to the previous team members, Dr. Jaejin Lee, Elena Winzeler, Amy Maxwell, and Bob Qui for inspiring my dissertation ideas. I also appreciate the friendship of the many graduate student colleagues, Yujung Ko, Mihyun Lim, and Emily Mckelroy.

Lastly, I greatly appreciate the support and encouragement of my family.

Examining Scientific Thinking Processes in Open-Ended Serious Games through Gameplay Data

by

Jina Kang, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Min Liu

Abstract: Research on scientific problem-solving emphasizes the importance of problem solving and scientific inquiry as central components of the twenty-first century skills. Research has shown that open-ended serious games can facilitate students' development of specific skills and improve learning performance through scientific problem-solving. However, understanding how students learn these complex skills in a game environment is a major challenge, as much research depends on typical paper-and-pencil assessments and self-reported surveys or other traditional observational and quantitative methods.

The participants of the study were 237 sixth graders from two middle schools in the Southwestern area of the United States. The students used an open-ended serious game called *Alien Rescue* as their science curriculum for three weeks. The purpose of this study is, first, to identify students' navigation behavior patterns in cognitive processes between at-risk and non-at-risk students within *Alien Rescue*. To accomplish this purpose, this study intends to use gameplay data by incorporating the integrated method of lag sequential analysis and sequential pattern mining together with a statistical

analysis. The findings confirmed that the integrated method helped to explore students' latent navigation behaviors as well as discover the differences of problem-solving processes between non-at-risk and at-risk students.

The second purpose of this study is to examine the relationship between students' learning performance and their scientific inquiry behaviors, which emerged as students engaged with Probe Design Center in this serious game. The results showed that the game metrics developed in Probe Design Center improved the predictions of both in-game and after-game performance. The cluster analyses with game metrics confirmed four unique groups regarding students' scientific inquiry behaviors in Probe Design Center. This study concluded that the integrated methods of serious games analytics enabled researchers to investigate in-depth cognitive processes and scientific inquiry behaviors within a specific cognitive tool, Probe Design Center, and discover unique behavior groups across different school settings. The researcher identified the challenges of at-risk students in their cognitive processes and highlighted the support needs for these students. Consequently, this study proposed an interactive dashboard using the data-driven evidences to provide teachers just-in-time information to support students' cognitive processes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

SIGNIFICANCE OF THE STUDY

The launch of *Sputnik 1*, the world's first artificial satellite, prompted policy makers to devise educational reform related to science curricula; that is, science literacy including both content knowledge and inquiry skills among academically diverse students (Barrow, 2006; Perkins, 1986; Stokes, 1997). Resultant post-Sputnik era instruction placed great importance on developing innovative science instruction that specifically emphasized science processes such as observation, classification, and inference (DeBoer, 1991). More recently, policy makers stressed the importance of twenty-first century skills—such as critical thinking and problem-solving—for academic or future employment success; this emphasis is reflected in various policy reform efforts (American Association for the Advancement of Science, 1993; National Research Council (NRC), 1997, 2012). Over the past few decades, policy reform documents show that scientific inquiry is another central skill for science literacy. The notion of scientific inquiry is considered to be a variety of activities involving scientific thinking and investigation such as making observation, examining any existing knowledge, interpreting data, and making predictions (NRC, 1996). Research on scientific problem-solving emphasizes students' need to possess both scientific inquiry as conceptual knowledge and problem-solving as procedural knowledge, both important components of twenty-first century skills (Clark-Midura, Dede, & Norton, 2011; Gott, Duggan, & Roberts, 2008; Lederman et al., 2014; Wecker et al., 2013). However, these complex skills can be difficult for students to learn (Duschl, Schweingruber, & Shouse, 2007; Hmelo-Silver & Azevedo, 2006). Lederman et al. (2014) recently asserted that a lack of research exists about improving students' understanding of scientific inquiry.

Expertise is as a key factor impacting students' scientific problem-solving processes and strategies (Jonassen, 2000; Wiley, 1998; Zimmerman, 2000). Researchers, therefore, have examined the components of expertise that indicate to teachers the individual differences in students' learning processes. However, research asserts that a great challenge in understanding students' learning processes is the use of traditional paper-and-pencil assessments such as pretest and posttest. Researchers specifically pointed out the poor alignment of traditional assessments with science content standards and the high cost of alternate assessments like hands-on performance tests (Clarke-Midura et al., 2011; Gobert, Sao Pedro, Raziuddin, & Baker, 2013).

Advanced technologies such as virtual worlds (e.g., Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007; Kamarainen et al., 2013), the intelligent tutoring system (e.g., Gobert, Kim, Sao Pedro, Kennedy, & Betts, 2015), and serious games (e.g., Sawyer & Rejeski, 2002) facilitate students' development of these twenty-first century skills involved in problem-solving and scientific inquiry. For example, a serious game embeds several attributes such as goals and challenges for students to learn certain skills, develop mastery, and enhance learning performance in the process of problem-solving (Foundation of American Scientists, 2006; Loh, Sheng, & Ifenthaler, 2015a). Researchers have employed traditional methodologies using traditional educational assessments (e.g., pre- and post-tests, self-reported surveys) to investigate the impact of serious games on learner engagement or the effectiveness on learner performance. However, recent research has raised some concerns that the use of traditional assessments is challenging to investigate learners' skill-building in serious games. In particular, students can take diverse ways of problem-solving in open-ended serious games (Squire, 2008). Due to the complex systems, students' cognitive processes are difficult to examine.

Serious games analytics has enabled researchers to examine learners' behaviors by tracking their learning processes (Loh, 2012; Wallner & Kriglstein, 2013). The use of *in situ* gameplay data—students' in-game actions traced *in situ*—helps researchers to investigate diverse student groups' (e.g., expertise, at-risk factor) in-game behaviors in a serious game. First, research on the use of technology have reported the difficulties students placed at-risk experience particularly with the complex nature of problem-solving (Darling-Hammond, Zielezinski, & Goldman, 2014; Samsonov, Pedersen, & Hill, 2006). Darling-Hammond et al. (2014) strengthened the key factors of computer-based learning environment for at-risk students such as interactive attributes, technologies for content creation and exploration, and teacher and peer supports. However, little is known about how at-risk students experience of learning complex skills in computer-based learning environments.

Second, several game metrics indicate different areas of expertise such as "time-to-task-completion rate, … mental representations of knowledge, … specific gaze patterns in scanning for information" (Loh & Sheng, 2014, p. 324). Additionally, researchers have devised new game metrics to measure these components of expertise within a serious game environment. According to Loh and Sheng (2015b), similarity/dis-similarity metrics measure different navigational sequences between novices and experts. However, there are limited empirical findings to suggest potential meanings in dynamic gameplay data and appropriate game metrics for specific learning behaviors (e.g., metrics for expertise) in a game context.

Since serious games analytics can provide more opportunities to measure, assess, or improve students' learning performance, researchers have developed a systemized analysis procedure that explains how to capture, analyze, and subsequently visualize a student's behavior (Loh, 2006; Romero, Ventura, Pechenizkiy, & Baker, 2010). Scholars

have attempted to develop learner behavioral profiles and measure learning performance via supervised or unsupervised learning techniques, types of machine learning. Supervised learning techniques are mainly used to develop a model with data from known labels to predict future data labels (data with unknown labels) such as predicting students' dropout based on their past activities in the learning management system. Two general categories in supervised learning techniques are classification and regression. Unsupervised learning is mainly used to glean hidden labels or group memberships from unlabeled data. The most common method is cluster analysis.

Diverse data mining techniques have been employed to serious games analytics. However, researchers have raised issues of the limitations of these techniques, which are not directed by theoretical principles in educational contexts (Clark, Martinez-Garza, Biswas, Luecht, & Sengupta, 2012; Gobert et al., 2015; Zhou, Xu, Nesbit, & Winne, 2010). There is still a lack of empirical studies of these techniques to inform educational pedagogy. In addition, little research has reported the relationship between learners' in-game behaviors and learning performance in serious games. Investigating in-game behaviors is critical to provide key evidences of different learning performance, which ultimately enhance different problem-solving strategies of students with diverse characteristics such as novice-to-expert and at-risk/non-at-risk.

Through the emerging technology of data visualization, researchers examine and visually present gameplay data to understand differences among individuals and demographic groups, discover patterns, and understand how these patterns relate to students' learning performance in a serious game context (Wallner & Kriglstein, 2015). A variety of visualization techniques address the challenges of interpretations derived from large amounts of data. For example, Liu, Kang, Lee, Winzeler, and Liu (2015) investigated tool usage patterns among different groups of students. The researchers used

action shapes to represent multivariate data using a variation of multiple parallel coordinates (Scarlatos & Scarlatos, 2010); they confirmed that visualization could reveal findings not easily detected using traditional methods. However, researchers are concerned about the complexities of representing high dimensional data, such as spatial-temporal data and difficulties in interpreting visualizations (e.g., node-link diagram) (Andrienko & Andrienko, 2008; Wallner & Kriglstein, 2013). Therefore, research combining traditional statistics with gameplay data analysis incorporating visualization techniques is needed to provide a holistic view of learners' behaviors and the relationship between behaviors and learning performance.

Essentially, scanty empirical research exists on students' scientific thinking processes via problem-solving and scientific inquiry skills usage within open-ended serious game environments. The use of traditional educational assessments and data mining techniques without consideration of educational theoretical principles is difficult to provide in-depth understanding of students' complex skills development in complex learning environments. Therefore, this study intended to explore different data mining techniques using *in situ* gameplay data in combination with *ex situ* data (i.e. science knowledge test) to investigate students' cognitive processes, identify in-game behaviors, and understand the relationship between students' in-game behaviors and learning performance.

This study is built upon extant studies using *Alien Rescue* to investigate students' cognitive process patterns. Prior studies on students' cognitive process patterns in *Alien Rescue* mostly employed statistical analysis with limited metrics such as frequency and duration of tool use. This study expands on previous research by incorporating the combination of statistical analysis with data mining and visualization techniques in an investigation of learning processes among diverse students to identify any meaningful

patterns. An exploration of different gameplay data analyses and visualizations to discover useful patterns with data mining techniques will enhance understanding of diverse learning methods in an open-ended serious game. This research will help inform the indicators of different students' scientific problem-solving to ensure success for all students.

PURPOSE OF THE STUDY

The purpose of this study is to understand how students solve a central problem through the examination of learning behavior patterns in an open-ended serious game using statistical analysis in combination with data mining techniques and visualization methods. Specifically, this study seeks to investigate sixth-grade students' problem-solving and scientific inquiry skills as cognitive processes of scientific thinking while they interact with various cognitive tools in an open-ended serious game designed for middle school science. This game is known as *Alien Rescue*. This study intends to employ both statistical methods and unsupervised learning techniques (e.g., sequential pattern mining, *k*-medoids cluster analysis) via a combination of *in situ* data (i.e. user-generated data from using the cognitive tools, solution texts) with *ex situ* data (i.e. science knowledge test).

Alien Rescue engages students in scientific investigations aimed at finding solutions to a complex problem (i.e. finding an appropriate home in our solar system for six alien species displaced from their home planets). *Alien Rescue* also provides a variety of cognitive tools to support students' problem-solving processes. Students with varying skill levels or differing characteristics can approach problems in various ways in *Alien Rescue*. Given the challenges of understanding learning processes in a serious game

environment, this study intends to employ different methods beyond statistical analysis to identify learning behavior patterns as captured by the students' *in situ* gameplay data. Two types of behavior patterns are expected to be identified in this study: navigation patterns derived from each student's sequence of different cognitive tool use and scientific inquiry patterns derived from students' Probe Design Center activities. Probe Design Center is one of the cognitive tools built into *Alien Rescue*, and is designed to support the scientific inquiry process by allowing students to generate and refine hypotheses and design their probes. Therefore, to investigate students' scientific inquiry patterns, five different scientific inquiry skills in Probe Design Center—defined as game metrics for measuring scientific inquiry skills in this research—were: (a) number of launched probes, (b) number of repeated trials, (c) amount of new information, (d) amount of redundant information, and (e) number of errors. Specifically, sequential pattern mining, lag sequential analysis, and *k*-medoids cluster analysis were performed to identify these patterns. Using statistical methods in combination with visualization techniques, this study then investigated students' behavior patterns and the relationship among patterns and learning performance.

RESEARCH QUESTIONS

This study seeks to investigate the following research questions:

1) Does the average posttest score significantly differ between at-risk and non-at-risk groups?
2) What differences exist between at-risk and non-at-risk students' navigational behaviors as they interact with various in-game tools?

3) What is the relationship between students' scientific inquiry behaviors in Probe Design Center and their learning performance?
4) What scientific inquiry behavior patterns emerge as students engage with Probe Design Center?
5) How can visualizations help to illustrate data-driven evidences of students' in-game behaviors to provide teachers just-in-time support?

TERM IDENTIFICATION

Problem-solving. Problem-solving refers to individuals' attempts to attain a goal for which the individuals may possess multiple solutions or no solution. In this study, problem-solving as a central cognitive process is considered as domain-general strategies required in scientific thinking and investigation.

Scientific inquiry. Scientific inquiry refers to the general practices that students investigate certain aspects of the natural world followed by subsequent activities such as observations, experimentations, or predictions along with scientific content and critical thinking. In this research, scientific inquiry as another integral cognitive process is considered as domain-specific knowledge required during scientific thinking and investigation.

Serious games. Serious games refer to games that are not for fun, enjoyment, or entertainment, but primarily a serious purpose. In contrast to commercial games created for entertainment use, serious games are meant to improve skills and learning performance through training and instruction (e.g., decision-making skills, combat performance).

Open-ended serious games. Open-ended serious game in this study is defined as serious games with multiple solution paths within complex functional spaces (Harpstead

et al., 2015). This study uses the concept of functional space wherein a game really takes place (Schell, 2008). A functional space is different from a physical space. Monopoly has a two-dimensional physical board (i.e. one paper board with a list of 40 real estate properties printed); however, it has only a one-dimensional space in terms of the function—a single line of 40 discrete points that links to each other in one loop.

Serious games analytics. Given the definition of serious games, researchers have concerns about how to measure improving skills and learning performance in serious games. Serious games analytics has been emerged to meet the needs of various stakeholders; discovering useful metrics for performance measurement, identifying significant predictors of expertise, designing better learning experience, and so forth. In short, serious games analytics refers to analytics or insights converted from gameplay data within a serious game for the purpose of performance measurement, assessment, or improvement.

Supervised learning. Supervised learning is one type of machine learning technique that is used for predicting future data labels such as high- or low-performing class. There are two broad categories of supervised learning techniques: classification when input data label is discrete; and regression when the label is continuous. Using both methods, a model learns from observations (i.e. input data) and improves its classification accuracy when more observations (i.e. new input data) are added, and then the model can make predictions of labels/classes of future data. The most common methods are decision trees, Bayesian networks, Linear Discriminant Analysis, K-nearest neighbor classifiers, and regression analysis.

Unsupervised learning. Another type of machine learning technique is unsupervised learning, which is mainly used for exploratory data analysis, in which researchers can find hidden patterns or group memberships from unlabeled data. The

most common method of unsupervised learning is cluster analysis, which is useful to separate learners into a certain number of groups/clusters when there are no predefined classification labels.

In situ data. *In situ* data are derived directly from learners' actions within the system. Typically, multiple parameters such as the number of clicks and duration of interaction are stored as logs *in situ*. These user-generated logs can then be used for understanding how an individual learner performs within a game environment, identifying any frequent navigation patterns across groups of learners, and visualizing the patterns

Ex situ data. *Ex situ* data are collected outside a system. The most common examples are user-surveys, pretest/posttest, talk-aloud, and interview. Both the user-surveys in a self-reported format and the pretest/posttest consider the game environment as a black box; that is, the data are collected only before or after learners interact with the game environment. Therefore, with *ex situ* data, researchers cannot easily assess how learners interact with the environment and this affects students' overall performance.

Chapter 2: Review of Literature

The purpose of the literature review is to provide theoretical foundation that guides this study including integral cognitive processes of scientific thinking such as problem-solving and scientific inquiry skills in serious games and the possibility and applicability of serious games analytics. The first section discusses scientific thinking through problem-solving and scientific inquiry skills. Relevant definitions of the problem-solving skills and research on expert-novice differences on the problem-solving process are discussed. Scientific inquiry skills and its historical background in conjunction with the policy reform documents are discussed. Both problem-solving as a procedural knowledge and scientific inquiry as a conceptual knowledge are discussed as important aspects of scientific investigation. The later section discusses the origins of open-ended serious games and serious games analytics. Advanced technologies including new game metrics adapted in serious games are reviewed in regard to tracing cognitive learning processes and measuring learning performance. Relevant research on at-risk and expertise are discussed. Lastly, different methods and visualization techniques applied to serious games analytics are discussed as integral to understand scientific thinking processes in an open-ended serious game.

COGNITIVE LEARNING PROCESSES

Problem-solving

What is a problem?

Jonassen (2004) defined a problem with two essential attributes; first, a problem as an unknown entity in some context representing the difference between a learner's

initial state and a goal state, and second, the learner perceives solving for the unknown as a worthwhile activity that has social, cultural, and intellectual value. Problems can vary in structuredness, complexity, dynamicity, and domain specificity or abstractness.

Specifically, problems can differ in how well structured they are—from well-structured problems to ill-structured problems (Jonnassen, 2004). Well-structured problems are often found in school systems, which demonstrate every component of the problem with a clearly defined initial and goal state and understandable solutions. Compared to a well-structured problem, ill-structured problems have unknown elements and have multiple solutions. There are no absolute or systematic criteria for assessing the solutions; therefore, learners need to describe their own thoughts or beliefs about the solution. Recent studies show differences in learning processes or performances between well-structured problems and ill-structured problems (e.g., Cho & Jonassen, 2002; Jonassen & Kwon, 2001; Schraw, Dunkle, & Bendixen, 1995).

In addition to the structuredness, problems vary in their complexity in terms of the related number of issues, functions, or variables, and dynamicity. For example, well-structured problems may have only a few variables, and in contrast, ill-structured problems possess many variables that can erratically interact with each other and in turn increase the difficulty of the problem (English, 1998). Well-structured problems are more stable, while ill-structured problems tend to be more dynamic (Jonnassen, 2004). For example, factors of a complex problem may keep changing over time.

Jonnassen (2004) lastly noted that problems are situated within a domain or context where individuals solve the problems differently relying on cognitive operations related to the specific domain or context. That is, individuals in different domains learn each form of reasoning skills related to the domain by solving the problems. Therefore,

problems can be described in terms of different levels of structuredness, complexity, and dynamicity within a specific domain.

Problem-solving

A problem exists in a situation, in which a learner attempts to reach some goals and to find out how to approach the goals (Chi & Glaser, 1985). Therefore, problem-solving refers to individuals' attempts to attain a goal for which the individuals may possess multiple solutions or no solution (Shunck, 2016). "Any goal-directed sequence of cognitive operations" (Anderson, 1980, p. 257) is referred to as problem-solving. Jonnassen (2004) described two critical aspects of the cognitive operations. First, problem-solving requires individual learners to build a mental model—known as the problem space. A mental model is composed of different kinds of knowledge: the structure of the problem, how to perform learning activities, and the appropriate use of procedures (De Kleer & Brown, 1981). Second, an individual actively manipulates and tests the mental model. Similarly, an information-processing model of problem-solving (Newell & Simon, 1972) includes a problem space, which consists of a beginning stage, a goal state, and solution paths. Learners construct a mental model of the problem and attempt to decrease a gap between the initial and goal states through the application of operations.

There are two historical perspectives on problem-solving: the trial-and-error approach, and intuitive knowledge or insight. Thorndike (1913) perceived problem-solving as trial-and-error and used cats' problem-solving ability to describe this process. Throndike found that the more that the cat tried, the less time the cat spent solving the problem; therefore, the cat learned through trial-and-error. There are many drawbacks to the trial-and-error approach such that it is often unreliable and not effective. For example,

repeated trials can waste time and not produce a plausible solution. Another perspective is involved in insight. Wallas (1926) formulated a four-stage model:

(1) Preparation: A time to learn about the problem and gather information that might be relevant to its solution.

(2) Incubation: A period of thinking about the problem, which may also include putting the problem aside for a time.

(3) Illumination: A period of insight when a potential solution suddenly comes into awareness.

(4) Verification: A time to test the proposed solution to ascertain whether it is correct. (cited in Shunck, 2006, p. 260)

Based on this four-stage model, Helie and Sun (2010) proposed a unified framework for understanding creative problem-solving including a more detailed conceptualization of the incubation and illumination stages. Although much research supports the existence of insight in problem-solving (e.g., Duncker, 1945; Durso, Rea, & Dayton, 1994; Kohler, 1925), there is a lack of research examining how learners develop and use insight, and more importantly, how teachers might implement an insight-learning framework in the classroom. Different types of problem-solving strategy have been investigated, with which learners can develop insight in problem-solving.

Problem-solving strategies

There are two types of problem-solving strategies: general strategies and specific strategies (Shunck, 2016). General strategies can be useful in different domains while specific strategies can be applicable to problems in a certain domain. Problem-solving strategies can either be general or specific; for example, analyzing subgoals in a given

problem can be useful in any domain. In addition, general strategies may not be useful with a familiar problem. There are two typical useful general strategies: generate-and-test and means-ends analysis.

According to Resnick (1985), the generate-and-test strategy can be used to test fewer solutions to show whether or not a learner achieves a goal. Learners use both prior knowledge to build the relative importance of all possible solutions, and current knowledge to select the most likely solution. In the means-ends analysis, an individual compares a goal and a current situation and identifies the difference between them to determine the best strategy for achieving the goal. Newell and Simon (1972) noted that people often used the means-ends-analysis to solve problems. Using the means-ends-analysis, they proposed a computer program, General Problem Solver (GPS), which was designed to provide essential processes that can be applied to solve various types of problems. Basically, the GPS algorithm assumes a goal is attained in a certain sequence, in which there are several subsequent goals for humans to attempt to achieve each. Then, the GPS transforms one into another and uses operations to eliminate the difference. There are two different types of means-ends analysis: working forward (i.e. from an initial state to a goal) and working backward (i.e. from a goal to an initial state).

Scientific Inquiry

There are some historical views of scientific inquiry. Dewey (1910) first introduced the notion of inquiry as a teaching strategy for K-12 science teachers. His strategy consisted of "sensing perplexing situations, clarifying the problem, formulating a tentative hypothesis, testing the hypothesis, revising with rigorous tests, and acting on the solution" (Barrow, 2006, p. 266). Dewey (1938/1977) also encouraged students to be

actively involved in learning by adding their personal knowledge of science and searching for answers. Later, Dewey (1944) modified his earlier interpretation of scientific methods by adding the concept of reflective thinking and suggesting these steps: "presentation of the problem, formation of a hypothesis, collecting data during the experiment, and formulation of a conclusion" (Barrow, 2006, p. 266).

The launching of Sputnik I in 1957 provided educators and policymakers the opportunity to examine the quality of science instruction in schools in the United States. As noted in Perkins (1986), the post-Sputnik era instruction put emphasis on science literacy including scientific knowledge, inquiry skills, and understanding of the nature of science. The National Science Foundation funded the development of an innovative science curricula that included biology, chemistry, physics, and earth science (e.g., Physics Science Study Committee, 1960). It specifically emphasized "thinking like a scientist" (DeBoer, 1991) and scientific reasoning processes such as observing, classifying, inferring, or controlling variables. Schwab (1966) also asserted that students should always consider science as a series of conceptual structures that can be modified where any new information or evidence is discovered.

As reported by Project Synthesis (Harms & Yager, 1981) including a review of the 1955-1975 literature and the 1977 national survey, most of the research on inquiry studied the content and strategy used by science teachers. Welch, Klopfer, Aikenhead, and Robinson (1981) described the reasons why teachers did not implement inquiry into classrooms due to the lack of materials, support, and teacher preparation. More recently, Anderson (2002) synthesized the literature on inquiry in science education, and he emphasized educators would integrate inquiry into their classrooms based on their beliefs and values about students and purposes of teaching. Anderson described the technical challenges (e.g., barriers presented by state assessments), political challenges (e.g.,

conflicts between science teachers or parents of how to teach science), and cultural challenges (e.g., different views of assessments).

The modern view of scientific inquiry combined with both historical (e.g., Dewey and Schwab) and recent perspectives were reflected in policy documents. The NRC (1996) described the notion of scientific inquiry as:

> a multifaceted activity that involves making observations; posing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results. Inquiry requires identification of assumptions, use of critical and logical thinking, and consideration of alternative explanations. (p. 23).

The NRC (1996) also identified six categories of inquiry to help students understand how and why scientific knowledge modifies and improves when new evidence, methods, or explanations occur in a scientific community:

1. conceptual principles and knowledge that guide scientific inquiries;
2. investigations undertaken for a wide variety of reasons—to discover new aspects, explain new phenomena, test conclusions of previous investigations, or test predictions of theories;
3. use of technology to enhance the gathering and analysis of data to result in greater accuracy and precision of the data;
4. use of mathematics and its tools and models for improving the questions, gathering data, constructing explanations, and communicating results;

5. scientific explanations that follow accepted criteria of logically consistent explanation, follow rules of evidence, are open to question and modification, and are based upon historical and current science knowledge; and

6. different types of investigations and results involving public communication within the science community. (To defend their results, scientists use logical arguments that identify connections between phenomena, previous investigations, and historical scientific knowledge; these reports must include clearly described procedures so other scientists can replicate or lead to future research) (as cited in Barrow, 2006, p. 270-271).

The Atlas of Scientific Literacy (AAAS, 2001) described scientific inquiry as comprising three categories: evidence and reasoning, scientific investigations, and scientific theories. In the book, *Inquiry and the National Science Education Standards*, the NRC (2000) described five essential features of inquiry for all grade levels including the following:

1. Learner engages in scientifically oriented questions; 2. Learner gives priority to evidence in responding to questions; 3. Learner formulates explanations from evidence; 4. Learner connects explanations to scientific knowledge; and 5. Learner communicates and justifies explanation (p. 29).

More recently, policymakers are emphatic about the need for 21st century skills. For example, the NRC's framework (2010) includes cognitive skills, interpersonal skills, and intrapersonal skills. The Partnership for 21st Century Skills' framework (2007) includes learning and innovation skills, life and career skills, and information, media and technology skills. Although the terms of skills are different in each framework, individual skills are similar across the frameworks (e.g., critical thinking, complex communication, problem-solving, self-regulation, and social skills).

There are many variations of inquiry learning and teaching in science education such as project-based science, problem-based learning, or model-based inquiry. According to Crawford (2014), all different cases should possess a central question that requires investigation and exploration. For example, problem-based learning (PBL) includes a complex real-world problem to encourage learning in science classrooms, in which students learn science concepts through understanding real-world problems, collecting scientific information they need, and reflecting on experiences in an active and collaborative learning environment. Regardless of the variations, the main facets of inquiry in science classrooms are to encourage students to learn scientific concepts as well as scientific explorations.

There have been many attempts to characterize scientific inquiry and integrate scientific inquiry with the concepts of particular domains of science over the past several decades, and this focus is reflected in a variety of policy reform documents. There is overall consensus that scientific inquiry refers to the general practices that students investigate certain aspects of the natural world followed by subsequent activities such as observations, experimentations, or predictions along with scientific content and critical thinking.

Inquiry and Assessment

While policy makers and educators paid strong attention to the notion of inquiry within the area of science teaching and learning for the last half of the 20th century, there were some challenges of implementing the inquiry instruction and learning in a classroom setting due to issues related to state assessments (Anderson, 2002; DeBoer et al., 2008). For example, as discussed in DeBoer et al. (2008), only eleven states provided

assessments strongly linked to content standards, and many educators claimed poorly written assessments did not properly align with the content standards.

There have been many attempts to address these challenges. In the 1990s, Maryland applied hands-on performance assessments in their science classrooms. Hands-on assessments help teachers to understand students' learning achievement (Clarke-Midura et al., 2011). However, as reported in many studies, students performed differently on similar tasks on various occasions, and these hands-on assessments are cost prohibitive and have limited validity compared to multiple-choice tests (Cronbach Linn, Brennan, & Haertel, 1997; Stecher & Klein, 1997).

Another attempt is the Project2061—a long-term initiative of the Advancement of Science (AAAS)—which has developed assessment items for middle- and early-high school science that are align with core ideas in national and state content standards and provide effective measures of students' understanding of science learning goals (AAAS Project 2061, n.d.). However, their tests are implemented using multiple-choice items, which are limited to standardized achievement tests that are traditionally implemented using paper-and-pencil formats and multiple-choice items. Specifically, multiple-choice tests are insufficient for reflecting complex science knowledge and understanding the nature of the students' critical thinking involved in the science inquiry process, but rather beneficial for determining a level of proficiency (Clarke-Midura et al., 2011). In sum, the limitations of these tests are due to the poor alignment with content standards, the cost of alternate assessments, and the insufficient format such as a paper-and-pencil format.

More recently computer-based environments have been designed to understand students' learning such as what they know and how they use their current knowledge and assess their process and strategies as an alternative way to the traditional assessments (e.g., Gobert et al., 2013; Quellmalz, Timms, & Schneider, 2009). These environments

can capture what an individual student is doing in the environments, that is, track the student's learning process *in situ* and use the captured data to assess their inquiry process and the final products they create, to measure their learning performance. A later section will discuss different data types involved in a computer-based environment, specifically in serious games.

Scientific Thinking through Problem-Solving and Inquiry

According to Zimmerman (2000), the general notion of scientific thinking or investigation relates to various activities such as "asking questions, hypothesizing, designing experiments, using apparatus, observing, measuring, predicting, recording and interpreting data, evaluating evidence, performing statistical calculations, making inferences, and formulating theories or models" (p. 102). Therefore, there have been attempts to view the scientific investigation as either conceptual (i.e. domain-specific knowledge) or procedural (i.e. domain-general strategies) aspects of scientific reasoning. In the domain-specific approach, learners investigate concepts using their current conceptual understanding and experiences with scientific phenomenon without conducting an experiment or investigating the results. On the other hand, the domain-general approach involves domain-general reasoning and problem-solving strategies; specifically, learners design an experiment and evaluate the findings from the experiment.

Research on scientific problem-solving involves the role of domain knowledge (Mayer, 2013), and all knowledge associated with scientific reasoning is either procedural or conceptual (Gott, Duggan, & Roberts, 2008; Gott & Murphy, 1987). Klahr and Dunbar (1988) emphasized the importance of both conceptual and procedural

knowledge and proposed an integrated model by incorporating domain-general strategies with domain-specific knowledge—known as the scientific discovery as dual search (SDDS) framework. SDSS as a cognitive process framework is based on the assumption that scientific discovery is a type of problem-solving where there are two problem spaces: hypothesis space and experiment space. SDSS includes three major components: searching the hypothesis space, searching the experiment space, and evaluating evidence. In this framework, learners can use prior knowledge to refine the search or they must carry out experiments in advance of developing an initial hypothesis. Wiley (1998) noted the organization of domain knowledge—how "accessible, proceduralized, integrated, and principled" (p. 716) the organization is—helps learners engage in problem-solving tasks. According to Mayer (2008), it is essential for a problem solver to recognize that the previous conceptual understanding of scientific phenomena is wrong and thereby needs to be changed, in other words, conceptual change is always involved in learning scientific problem-solving. For example, Chen and Klahr (1999) noted that direct instruction in designing controlled experiments—known as the Control of Variables Strategy (CVS)—can help learners engage in the scientific reasoning process of testing hypotheses. Using this strategy, a learner changes only one variable in the experiment to see the effects of the variable. Therefore, the learner should be able to understand the procedures and concepts of the controlled experiment.

Gobert et al. (2015) noted that students have difficulties when designing controlled experiments. Gobert and her colleagues observed that students might collect limited evidence to test their own hypotheses, attempt only one trial or repeated trials with the same condition (e.g., same variables), or revise too many variables. Students have difficulties with complex systems such as with multiple independent variables that interact with each other. In that case, students need to be careful about which variable

they change, since results can be influenced by the interaction. Hmelo-Silver and Azevedo (2006) asserted a challenge of measuring younger students' inquiry strategies, specifically in middle schools, since they have more difficulties to understand these complexities of system and scientific inquiry process. Lederman et al. (2014) asserted a lack of research on improving students' understandings of scientific inquiry in K-12 science education. They specifically emphasized that conducting inquiry would not result in developing scientific inquiry process knowledge. In addition, since scientific inquiry especially involves critical and logical thinking, traditional assessments such as science achievement tests do not demonstrate students' conceptual knowledge and procedural knowledge related to inquiry (Clarke-Midura et al., 2011; Gobert et al., 2013; NRC, 1996; Quellmalz et al., 2009).

Summary

Scientific problem-solving research shows that learning to solve scientific problems involves a conceptual change only achieved when students possess both conceptual and procedural knowledge—central components of twenty-first century skills. However, students often face challenges in conducting inquiry within complex systems and understanding the processes involved in scientific inquiry. Extant research also highlights the limitation of typical school assessments' ability to measure student inquiry strategies.

Overall, the policy reform documents over the past century acknowledge that problem-solving and scientific inquiry skills—central cognitive processes during learning. Although different perspectives exist regarding scientific thinking or investigation, the consensus holds that all knowledge associated with scientific thinking

is either procedural or conceptual. Specifically, in science literacy education, problem-solving skills as domain-general strategies are integrated with scientific inquiry skills as domain-specific knowledge, which together support students' scientific thinking and investigation.

Comprehension of the processes involved in scientific problem-solving and inquiry is a central, albeit difficult skill for students to learn and for educators to assess. Research highlights the importance expertise has on students' scientific problem-solving processes and strategies; findings provide teachers with specific advice about how different students can develop their own strategies. However, also research documents the difficulty in assessing these processes via traditional educational assessments.

There has been research interest in the use of advanced technology such as simulations, 3D virtual learning environments, or serious games to support students' scientific thinking and assess students' scientific inquiry process skills, which will be discussed in the next section.

SERIOUS GAMES ANALYTICS

Educational Games to Serious Games

During the past decades, different terms of digital games for education have emerged such as digital game-based learning and serious games. *The Oregon Trail* game was the first popular learning tool to teach students about the realities of pioneer life in schools between the mid-1980s and mid-2000s. The game was designed for challenging learners in a fun way, but not assessing performance. During that period, educators also began to develop their own games using emerging authoring tools such as Flash or

Authorware (Loh et al., 2015a). The purpose of computer-based instruction is teaching specific skills or complex concepts in a certain domain. Similar to *The Oregon Trail*, these early educational games do not assess students' learning performances. However, these games led people to become interested in edutainment—combining education and entertainment. As a result, it contributed to enhancing students' engagement by adding fun and motivating features to traditional learning materials. More recently, the term, digital game-based learning (DGBL), became popular by the renowned works of Prensky (2001) and Gee (2003). Prensky noted the emergence of DGBL and argued that students in the last decades of the 20th century—surrounded by technologies, defined as digital natives—were different from their ancestors in terms of how to think and process information. Therefore, implementing DGBL as a learning tool in a classroom can address the gap. Gee (2003) considered a video game as a good teaching tool that can foster creative thinking and identified the list of learning principles commonly found in a good game (e.g., identity development, well-ordered problems, performance encouragement before competence) and discussed how these games affect how people learn.

In the early 2000's, there were two major incidents that eventually led to greater interest in using serious games for learning: a detailed report on improving public policy through game-based learning and simulation (Sawyer & Rejeski, 2002) and a first well-known serious game, *America's Army*. The term "Serious Games" first appeared in the book "Serious Games," and in contrast to entertainment, "these games have an explicit and carefully thought-out educational purpose;" however, Abt added, "this does not mean that serious games are not, or should not be, entertaining" (Abt, 1970, p. 9). More recently, several researchers attempted to define serious games as games that are not for

fun, enjoyment, or entertainment, but primarily a serious purpose (Michael & Chen, 2006; Sawyer, 2009; Zyda, 2005).

Loh et al. (2015a) noted that the definition of serious games—that is, any digital games not for entertainment—is way too broad. Instead, they suggested considering serious games with different characteristics such as the essential game attributes identified in the National Summit on Educational Games (Foundation of American Scientists, 2006): for example, "clear goals, repeatable tasks (to build mastery), … , encouraging increased time on task (through motivation)," (p. 8-9). Along with these attributes, learners should be able to learn and improve different skills and the processes such as problem-solving or decision-making.

In sum, although serious games can have an entertaining element, they are primarily designed for learning and training a variety of audiences (e.g., professionals, consumers, students) for real-world situations (e.g., Djaouti, Alvarez, Jessel, & Rampnoux, 2011; Zyda, 2005). The field of serious games has grown and involved with various domains and subject areas; therefore, different game structures have been considered to cover complex learning goals such as fostering problem-solving skills (Harpstead et al., 2015). The following section explores specific attempts of conceptualizing serious games with open-ended environments.

Open-Ended Serious Games

There have been different views on the notion of open-ended serious games. Squire (2008) identified "open-ended games" as possessing a strong potential for developing students' problem-solving, productive and digital technologies literacy.

Typically, learning occurs throughout the process of understanding the game system, conducting experiments, and communicating with other learners. Squire proposed a framework that distinguishes between game genres using several key variables such as time to completion, timescale, open-endedness, and modes of creative expression. For example, targeted (e.g., puzzles) or linear games (e.g., Ninja Gaiden) take a few hours to a month to complete with low open-endedness. Open-ended games can take a few months to years with high open-endedness. Squire identified two different types of open-ended games: massively multiplayer online (MMO) games (i.e. persistent worlds) and open-ended simulation games. For instance, both *River City* (Nelson et al., 2007) and *Quest Atlantis* (Barab et al., 1999) are examples of persistent worlds that include key features of MMOs such as multiple avatars and a 3D virtual world that invites exploration along with interactive features. The main feature of another type, open-ended simulation games, is multiple solution paths, where each learner develops their own learning spaces for knowledge creation or discovery.

Similarly, Spring and Pellegrino (2011) noted that open games typically have multiple pathways in a less guided learning environment. Their major learning goals are conceptual learning and science process skills, which involves a variety of processes such as learning through practice or failure (Gee, 2003; Squire, 2008). For example, learners are allowed to test multiple solution paths throughout the game although some penalties might be incurred. As a result, this complexity enhances learners' curiosity and replayability, but also challenges designers and researchers to assess their learning.

Harpstead et al. (2015) considered a game structure as a primary key feature of open-ended games. They defined open-ended games using the concept of functional game spaces wherein a game actually takes place (Schell, 2008). Schell (2008) used several games (e.g., *Monopoly*, *Twenty questions*) to describe the concept. Monopoly has

a two-dimensional physical board (i.e. one paper board with a list of 40 real estate properties printed); however, it has only a one-dimensional space in terms of the function—a single line of 40 discrete points that links to each other in one loop. *Twenty questions* is a game in which one player can ask twenty questions to guess what another player is thinking in mind. There is no physical space in the game; however, Schell suggested this game has three spaces as "Mind of the answerer," "Conversation space," and "Mind of the questioner" (p. 135, see Figure 1). From this perspective, the functional solution space in a game determines its open-endedness (Harpstead et al., 2015). If a game has a simple functional space (e.g., *Monopoly*), it is less open-ended. In an open-ended serious game with complex solution spaces, it can be challenging to understand learner behavior in the game and guide learners even minimally how to approach a goal. Schell (2008) asserted that taking these things into consideration is essential for a game designer.

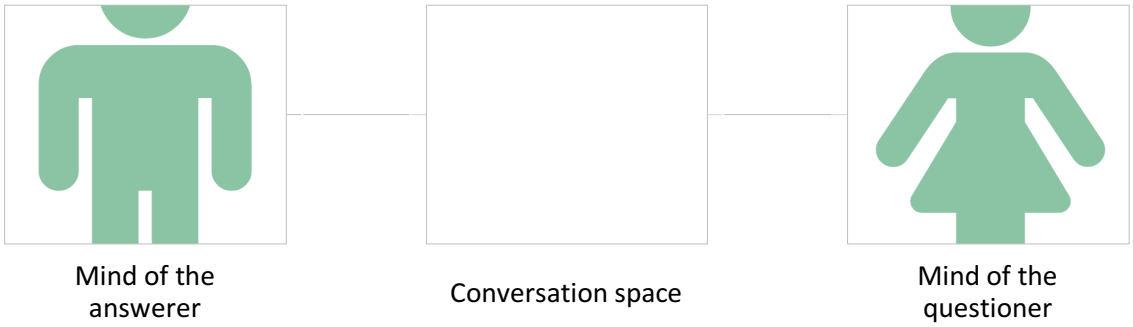


Figure 1: Game space in "Twenty questions"

Combining these prior attempts of defining an open-ended serious game, this study defines an open-ended serious game as a serious game with multiple solution paths within complex solution spaces. In the following section, serious games analytics will be

discussed to address the challenge of understanding learner behaviors in an open-ended serious game.

Serious Games Analytics

As the field of serious games has grown, researchers have paid more attention to the area of serious games analytics. Loh et al. (2015a) described serious games analytics as:

> actionable metrics developed through problem definition in training/learning scenarios and the application of statistical models, metrics, and analysis for skills and human performance improvement and assessment, using serious games as primary tools for training (p. 23).

The major interests of researchers are to understand what learners do in serious games and investigate the effectiveness of games by tracing user-generated data. Therefore, researchers can inform educators and developers of these insights to support better learning design and improve skills and performance of students.

Serious games analytics has created possibilities of tracing users' behaviors in a game beyond the traditional performance assessment (Loh, 2012; Wallner & Kriglstein, 2013). Recent research has focused on users' behavior captured within the game environment *in-situ* as evidences of users' learning performance (Loh & Sheng, 2014; Schmidt & Lee, 2011). Traditional data methods such as surveys or pretest and posttest study designs often cannot capture a user's intermediate learning process or changes in learning performance (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013; Gobert et al.,

2013; Loh et al., 2015a). Serious games analytics, on the other hand, can capture users' learning comprehension while highlighting their learning process and performance improvement (van Barneveld, Arnold, & Campbell, 2012). To address these issues, different types of user-generated data driven from serious games and types of metrics that can be used to measure learning in a game will be discussed in the following sections.

User-Generated Data: Ex Situ Data vs. In Situ Data

Two different types of user-generated data—*ex situ* and *in situ* data—measure what learners do in a game-based learning environment and can be used to assess performance. The most common examples of *ex situ* data are user-surveys, pretest/posttest, talk-aloud, and interview, which are collected "outside the system" (Loh et al., 2015a, p. 16). Both the user-surveys in a self-reported format (Fan et al., 2006) and the pretest/posttest consider the game environment as a black box; that is, the data are collected only before or after learners interact with the game environment. Therefore, with *ex situ* data, researchers cannot easily assess how learners interact with the environment and this affects students' overall performance. Despite the limitations of *ex situ* data, recent research has shown that user-surveys, pretest/posttest, and questionnaires are the most prevalent techniques (Bellotti et al., 2013; Smith, Blackmore, & Nesbitt, 2015). Loh and Sheng (2015a) noted that researchers who lack programming skill favor a qualitative approach such as talking aloud, focus group interviews, or video recordings during the game sessions.

In comparison, *in situ* data are derived directly from learners' actions within the system. Typically, multiple parameters such as the number of clicks and duration of interaction are stored as logs *in situ*. These user-generated logs can then be used for

understanding how an individual learner performs within the environment, identifying any frequent navigation patterns across groups of learners, and visualizing the patterns (Scarlatos & Scarlatos, 2010, see the details in the visualization section). Limitations of *in situ* data include the inability of providing context such as why a learner is doing something or if a learner is having fun or not. (Wallner & Kriglstein, 2013). Therefore, scholars have attempted to identify potential meanings for the parameters. Linek, Öttl, and Albert. (2010) emphasized that each parameter can be considered a specific behavior indicator which includes subjective meaning. As user gameplay data becomes a prevalent feature in serious game environments (Loh & Sheng, 2014), serious games analytics can track users' decision-making processes to further support personalization of instruction (Linek et al., 2010; Liu et al., 2015; Loh, 2012; Reese, Tabachnick, & Kosko, 2015). Therefore, in comparison to *ex situ* data, with *in situ* data researchers can consider the game environment as a white box (Loh et al., 2015a). That is, *in situ* data, which is collected without interrupting the learners, collects information on such as how many and how fast a task is completed. Compared to *ex situ* data obtained by human data-input, *in situ* data are less subjective and erroneous (Fan et al., 2006; Loh et al., 2015a; Quellmalz et al., 2009).

One of the most prevalent techniques of *in situ* data collection is *telemetry*. *Telemetry* has been used in various fields such as computer science, ecology, or biology—with which any computer or mobile applications transmit user-generated data to a remote server for storage or further analysis. The logs generated during this process—usually in a plain text format—are available to use for the analysis after preprocessing the logs. In addition, an online database combined with *telemetry* enables real-time analytics, which allows researchers to collect gameplay data remotely, and

instead, the system transmits user-generated data to a server for storage and analysis (Loh, 2006; Loh & Sheng, 2015a).

Game Metrics

How to collect *in situ* user-generated data (i.e. gameplay data) is a great concern for researchers and designers in the serious games analytics field. Researchers can interpret learners' repeated actions as behaviors, and therefore, they can discover a pattern or trend (Loh et al., 2015a; Wallner & Kriglstein, 2013). It is also essential for researchers to understand what type of learner actions or behaviors can lead to better learning performance.

Metrics that could appropriately measure learner performance would vary depending on the purpose of the game. There are some general metrics that can be useful regardless of the differences in purpose; for example, *time to completion* is one of the prevalent performance metrics in serious games research (Canossa & Drachen, 2009; Loh & Sheng, 2015a). Many game-based learning environments apply *time to completion* as a criterion for evaluating learning performance. According to Loh and Sheng (2015a), the *time to completion* metric can be interpreted differently depending on the situation; for example, fast speed can be considered as an indicator of a positive learning outcome or a negative learning behavior (e.g., rash decision). Therefore, appropriate game metrics in a variety of situations is needed to evaluate learners' skills or performance improvement.

Research on At-risk Students

Since U.S. Department of Education (1983) first defined the term *at-risk,* researchers used the term to describe students with various characteristics (Hammons-Bryner, 1995). For example, at-risk students can be defined as students who show a lack

of interest, motivation, or self-direction in learning; negative attitude toward teachers; boredom at schools (Ponticell, 2001). There is a growing number of at-risk students (Barr & Parrett, 2001). Throughout educational reforms that are concentrating strategies to close academic achievement gap, educators and policymakers have sought solutions regarding the uses of innovative technologies for at-risk students (Darling-Hammond, Zielezinski, & Goldman, 2014).

Darling-Hammond et al. (2014) highlighted three primary factors of success in learning for at-risk students who are particularly learning new skills in a technology-enhanced environment: interactive attributes, technologies for content creation and exploration, and teacher and peer supports. Research on interactive learning found that at-risk students learned quadratic functions using an interactive learning environment significantly more than other students in a traditional classroom setting using lecture, note taking, drill, and practice (Bos, 2007). This study highlighted the needs of developing an interactive environment that involves all levels of students in higher-order thinking skills. A number of studies have found that students show stronger engagement and skill development when they work with teachers and interact with other students. Kim and Lee (2011) evaluated online learning satisfaction of underprivileged students and illustrated the needs of teacher assistance in online learning. Particularly, students reported the just-in-time support and encouragement from the teachers played a critical role in increasing their academic standing.

Researchers have raised concerns over the effectiveness of computer-based learning particularly with student-centered pedagogy for students placed at-risk. Samsonov et al. (2006) examined the effectiveness of computer-based PBL with at-risk students and identified the at-risk factors related to performance in the PBL environment. The results revealed that most of at-risk students showed the feeling of boredom at the

beginning of the activity due to a lack of structure and metacognition, or confusion. The authors also highlighted prior academic performance as a critical at-risk factor in explaining different performance in the environment. For example, students with above average prior academic performance became more strategic in their problem-solving compared with the students with lower performance. However, the authors suggested that a computer-based PBL can be effective for students who are lower average academically with peer support; that is, collaboration with higher performing students.

A student-centered approach in PBL enhances students' thinking skills and self-direction, and a well-designed PBL task enables students to acquire inquiry and reasoning skills. To positively effect on students' knowledge building, the learning environment must provide structured information that can track students' problem-solving processes at any times (Gijbels, Dochy, Van den Bossche, & Segers, 2005; Hmelo-Silver, 2004) However, little is known about the at-risk students' learning processes or challenges in computer-based environments that requires higher-order thinking skills such as reasoning through problem-solving. In addition, the majority of studies largely depend on students' self-reported surveys, pre-and post-tests, or observations, which are limited in their understanding of how at-risk students use a computer-based learning environment such as serious games. Many existing studies on learning analytics were limited to use a frequency or duration as a metric to measure learners' behaviors within an environment. In addition to limited data sources and metrics, the majority of research used typical statistical methods such as ANOVA or descriptive analysis, which cannot account for in-depth learning processes of individual learners in computer-based environments. To help close this gap, the concepts of serious games analytics must be applied, wherein appropriate features/metrics can be determined and various techniques beyond traditional

statistical methods can be employed to better understand how differently diverse learners play a serious game.

Research on Expertise

The game metrics of particular interest to this study are metrics for expertise. Research on the behavioral and cognitive differences of individuals such as experts, skilled individuals, and novices have been well studied since the twentieth century (e.g., Bryan & Harter, 1899). As discussed earlier, the difference between experts and novices during problem-solving processes is a well-known phenomenon (Dreyfus, 2004; Dreyfus & Dreyfus, 2005; Jonassen, 2000; van Merriënboer, 2013; Wiley, 1998). Learners can overcome their limited working memory by consciously practicing a given task and eventually improving their expertise over time. For example, novices first tend to follow rules without thinking carefully; eventually, they learn how to apply rules correctly over time as well as attain more competency. Rule application and competency are a measurable change found in learners' action sequences during the problem-solving process (Dreyfus, 2004). Dreyfus determined five ordered stages of skill acquisition: novice, competence, proficiency, expertise, and mastery.

Several well-known indicators of expert-novice performance differences exist: "time-to-task-completion, mental representations, dynamic decision-making, gaze patterns, neural/perceptual responses" (Loh, Sheng, & Li, 2015b, p. 148). These have been proven with different learner behaviors through observance of various learner actions or errors (e.g., frequency, types) over a certain time period. In a recent study, researchers considered learners' action sequences (e.g., navigational sequence {Place A, Place B, Place C, Place A, Place F}) to be an important indicator for performance measurement and suggested various similarity/dis-similarity metrics (e.g., Jaccard

coefficient, Loh & Sheng, 2014) to view how similar or dissimilar novices' action sequences are from experts. For example, two players' action sequences can be compared using a Jaccard coefficient. A Jaccard coefficient 1 indicates that the two players' sequences are identical—that is, if one player is an expert, another player is "likely-experts" (p. 149); a coefficient 0 indicates that the two sets are totally different—that is, another player is a novice.

Game designers and researchers should consider using these attempts to devise new metrics when measuring a specific learning skill required in a certain scenario. Therefore, comprehending the nature of *in situ* data, the diverse types of game metrics, and the proper use of each to assess learning outcomes will facilitate better learning performance. The following section examines a number of techniques suitable for serious games analytics, such as profiling learner behaviors and measuring learning performance.

Methods towards Serious Games Analytics

There are different types of analytics such as Game Analytics, Learning Analytics, and Serious Games Analytics as to their primary purposes. Game Analytics is meant to develop monetization strategies by improving game design (e.g., Seif El-Nasr, Drachen, & Canossa, 2013). Learning Analytics is purposed to provide dynamic educational information to optimize learning and the environments such as Learning Management System and Intelligent Tutoring System (e.g., Siemens, 2013), in which serious games might be included. Serious Games Analytics supports knowledge acquisition or skill development (e.g., Bellotti et al., 2013). There are general metrics that can be used across these fields such as *time of completion*; however, it is worth noting

that different sets of metrics should be considered and developed for achieving the goal of each field (Loh et al., 2015a).

Since Serious Games Analytics and Learning Analytics are still new fields, both groups mostly use methods commonly found in the field of Game Analytics. Obviously, there is a similar intention to understand users (i.e. learners or game players) among Game Analytics, Serious Games Analytics, and Learning Analytics groups; that is, the groups share the idea of classification such as classifying users' knowledge, motivation, or behavior (Hämäläinen & Vinni, 2010). Specifically, researchers in the Serious Games field can develop learner behavioral profiles or assess learning performance using supervised or unsupervised learning techniques.

First, unsupervised learning techniques are mainly used for exploratory data analysis, in which researchers can find hidden patterns or group memberships. The most common method of unsupervised learning is cluster analysis, which is useful to separate learners into a certain number of groups/clusters when there are no predefined classification labels. The groups are determined by measuring similarity defined by a metric such as Euclidian or probabilistic distance. There are several common clustering algorithms such as Hierarchical clustering, k-Means clustering, and Gaussian mixture models. It is an exploratory process first to identify the best solution for a given task (including from selecting a number of clusters or an appropriate clustering algorithm to evaluating the result), and second, to interpret the result using domain knowledge. Therefore, it is essential to understand the strengths and weaknesses of each approach to obtain meaningful patterns from user-generated data (Drachen, Thurau, Togelius, Yannakakis, & Bauckhage, 2013).

Supervised learning techniques are used for predicting future data labels such as high- or low-performing class. There are two broad categories of supervised learning

techniques: classification when input data label is discrete; and regression when the label is continuous. Using both methods, a model learns from observations (i.e. input data) and improves its classification accuracy when more observations (i.e. new input data) are added, and then the model can make predictions of labels/classes of future data. The most common methods are decision trees, Bayesian networks, Linear Discriminant Analysis, K-nearest neighbor classifiers, and regression analysis. Once a classification method is selected, researchers need to prepare for a sample dataset with known labels and divide the dataset into two sub-sets—a training set and a test set. First, the classifier is run with the training set and then tested with the test set, which their classes are hidden, to see how accurately the classifier classifies the cases in the test set. If the classification accuracy is too low, we can either modify the data to search for a better model, change the algorithm, or switch to another classification method. Hämäläinen and Vinni (2010) compared different classification methods using several criteria such as a form of class boundaries (i.e. linear, non-linear), accuracy on small data sets, working with incomplete data, supporting mixed variables (e.g., numeric, categorical), and computational efficiency (p. 70, see Table 1).

	Decision Trees	Naïve Bayes	General Bayesian	FFNN	KNN	SVM	Linear regression
Nonlinear boundaries	+[a]	(+)	+	+	+	+	-
Small datasets	-	+	+/-	-	-	+	+
Incomplete data	-	+	+	+	+	-	-
Mixed variables	+	+	+	-	+	-	-
Natural interpretation	+	+	+	-	(+)	-	+
Efficient reasoning	+	+	+	+	-	+	+
Efficient learning	+/-	+	-	-	+/-	+	+
Efficient updating	-	+	+	+	+	-	+

Note. [a] "+" indicates the method supports the property, - that it does not. (FFNN: Feed-forward neural network, SVM: Support vector machine, KNN: K-nearest neighbor)

Table 1: Comparison of Classification Methods

Serious games enable researchers to understand in-depth learners' behaviors by tracking the information of locations and times of learners' actions (Loh & Sheng, 2015a). However, not all methods can be used for understanding spatial-temporal user behaviors. One of the most common methods, Bayesian Network, cannot measure

spatial-temporal nature of the data; therefore, it is not possible to track when or where individual users complete a specific goal within a game. Therefore, it is essential to seek for a more appropriate method that can account for spatial-temporal data for serious game analytics.

There are many attempts to deal with spatial-temporal data. Map & Analyze Patterns & Structures Across Time (MAPSAT, Frick, Myers, Thompson, & York, 2008) can analyze temporal or structural patterns/relations of educational data instead of mathematical relations (i.e. linear function). There are two approaches in MAPSAT: Analysis of Patterns in Time (APT) that maps temporal relations (e.g., *Event A* proceeds *Event B*, *Event A* co-occurs with *Event B*) and Analysis of Patterns in Configuration (APC) that maps structural relations (e.g., *Event A* affects-relation *Event B*) (Frick et al, 2011). Myers and Frick (2015) proposed APT could be used to assess the learning trajectory of an individual learner within a serious game. MAPSAT measures temporal and structural patterns by observing empirical phenomena; however, it does not provide statistical significance of a specific pattern, but instead conditional probabilities of patterns. Therefore, a linear model approach can be employed in addition to the patterns identified by MAPSAT for the generalization purpose.

Sequential pattern mining (Agrawal & Srikant, 1995) is another technique to examine students' sequential behavior patterns in a computer-based learning environment. Sequential pattern mining was first introduced to identify customer purchase sequences from a large database of customer transactions. This discovers a list of frequent sequences with a certain condition that the occurrence of the sequences must be greater than a certain user-specific *minimum support*. For example, researchers can specify a certain percentage of total students need to support a frequent sequential pattern. Researchers have raised some concerns of sequential pattern mining (Zhou et al.,

2010). Researchers might not be interested in certain actions; therefore, they need to filter out non-meaningful actions recorded in logged data. In general, computer-based learning environments record students' every single action including mouse clicks generated by students' inexperienced keyboard or mouse use. Therefore, translating raw logged data to meaningful actions is critical to extract relevant behavior patterns. In addition, a sequential analysis is unable to identify an exact timing of a sequence (Clark, Martinez-Garza, Biswas, Luecht, & Sengupta, 2012). Thus, seeking for an appropriate way of analyzing spatial-temporal data is a critical factor to deal with diverse researchers' interests and concerns.

Another issue of pattern mining is ignoring quantity information included in mined patterns that can provide insights to user behavior (Kim, Lim, Ng, & Shim, 2007). For example, a basic sequential pattern is a $<s_1, s_2, s_3, \ldots s_m>$, where $s_j = \{i_{j,1}, \ldots i_{j,nj}\}$ is an itemset. Kim and his colleagues pointed out that these sequential patterns only show their qualitative nature without quantitative information of each item $i_{j,k}$, while actual applications record quantitative information in their logs. For example, the sequential pattern, ([pants, 3]), ([jacket, 2], [sweater, 4])—often found in any marketing datasets—shows customers frequently purchase three pants first and two jackets and four sweaters later together. Therefore, the researchers proposed the techniques called SQUIRE (Sequential pattern mining with quantities) to identify quantitative sequential patterns. However, this technique is not developed for educational applications. As asserted by Zhou et al. (2010), when employing the sequential pattern analysis algorithms, researchers need to apply domain knowledge to find an appropriate way to filter out numerous meaningless patterns.

Lag sequential analysis (LSA)—sequential hypothesis testing—is based on sequential data, in which researchers assume continuity between items or actions. Once

transitional frequencies between items in a contingency table are calculated, standard statistical techniques can be used to determine certain transitions significantly occurred more often than others (Bakeman & Quera, 2011). More specifically, LSA yields a series of sequential analysis matrix calculations: transitional frequency matrix, transitional probability matrix, and adjusted residuals table (Bakeman & Gottman, 1997). Adjusted residuals (i.e., z-scores) of each transition were calculated to determine if the transitional probabilities deviated significantly from the expected value. For instance, a z-score above 1.96 denotes the behavioral transition from an action to another action in a certain group of users reaches a significant level of 0.05 ($p < .05$), that is, the transition occurs at the frequency greater than chance.

LSA has shown to be useful for understanding human computer interaction behaviors (Chung & Baker, 2003; Hou, 2015; Pohl, Wallner, & Kriglstein, 2016). Through sequential structure of users' interaction with computers, their use of technology and adaptation to various situations within a system such as how users solve a problem can be revealed (Sanderson & Fisher, 1994). Research on interaction processes with visualization systems found interactions patterns and cognitive processes that occur with a higher probability than others by conducting LSA (Pohl et al., 2016). The authors indicated the findings of interaction processes can be used to make inferences about users' reasoning processes. Chung and Baker (2003) performed LSA using users' logged actions in an interactive learning environment and found these sequential actions can be used as a measure of problem-solving processes.

Serious games, in contrast to commercial games, are meant to improve skills and learning performance. Serious game can be used as a tool for investigating learning behavior and measuring performance with *in situ* data. The recent research highlighted the importance of understanding the nature of *in situ* data and devising new game metrics

to understand individual differences in learning behavior and identify components of expertise in-game to facilitate better learning performance. However, there is a lack of research concerning game metrics that are applicable in different scenarios, especially, in the educational domain. There have been several attempts to identify spatial and temporal learning trajectories by using various data mining techniques, however, there is a strong need to apply domain knowledge to gameplay data and interpret results. Given the difficulties of understanding dynamic gameplay data, various graphical representation techniques have emerged to support researchers in understanding dynamic data and interpreting the complex patterns in the educational domain.

Visualization Techniques

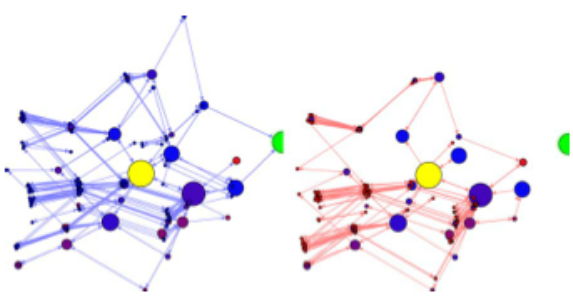
Various graphical representation methods have been used to assist serious games analytics in the recent literature. There have been several attempts to find standardized analysis procedures to track learning processes and then visualize the results (Loh, 2006, Romero et al., 2010; Romero & Ventura, 2013). Wallner and Kriglstein (2013) reviewed several techniques and tools to visualize the large amount of multi-dimensional temporal-spatial data to understand learner behavior in a game context (Wallner & Kriglstein, 2013). More recently, Wallner and Kriglstein (2015) discussed the benefits of visualizations for various stakeholders. Specifically, researchers can assess game design and pedagogical effectiveness, teachers can provide just-in-time feedback by monitoring students' progress, and learners can monitor their learning progress for self-reflection and collaboration with others.

Scarlatos and Scarlatos (2010) identified multiple representation techniques such as glyph-based techniques and parallel coordinates to support systematic analysis to

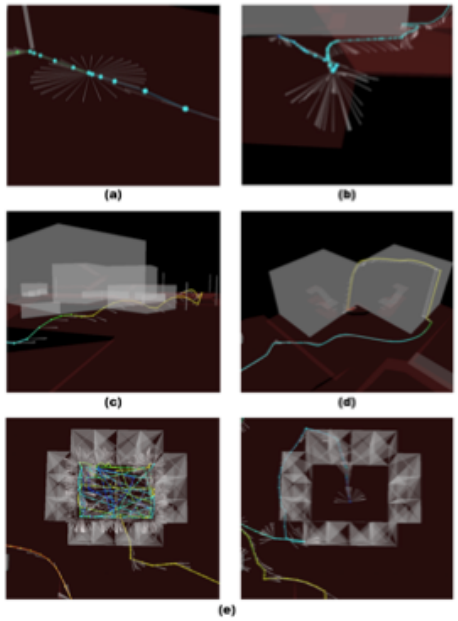
discover behavioral patterns. Wallner and Kriglstein (2013) identified five categories of the most common representation techniques: chart and diagram, heatmap, movement visualization, self-organizing map, and node-link approach.

The most prevalent representation techniques are charts and diagrams, with which researchers can represent simple game metrics such as individual students' time to completion rates and the number of completed tasks. Scarlatos and Scarlatos (2010) further applied the concepts of charts and diagrams to visualize different learners' behaviors, in which they proposed the idea of action shapes to represent multivariate data using a variation of multiple parallel coordinates. The results show behavioral differences between different groups of learners such as experts versus novices. A heatmap uses a color gradient to visualize spatial learner behaviors. For example, heatmaps can be used to visualize how long an individual user stayed at a certain location in a two-dimensional space. Drachen and Canossa (2009) further used the concept of heatmaps by overlaying multiple layers of different behaviors.

There are also attempts to represent high-dimensional data including both spatial and temporal information through node-link diagrams and movement visualizations. Using node-link representations, multiple variables from gameplay data can be mapped to different components of the diagram and then projected onto a two-dimensional space. Since movement visualization can provide a detailed learning path, it is often used to test usability during the game development phase. However, some pitfalls of a node-link diagram and movement representations have been reported such that a number of states or nodes might provide a cluttered visualization (Wallner & Kriglstein, 2013; G. Andrienko & N. Andrienko, 2010) (see the example in Figure 2).



(a) Winners (left) and losers (right) at the same game state. (p. 3, Anderson, Liu, Apter, Boucher-Genesse, & Popović, 2010).



(b) Six different player behavior patterns: (a) pirouette (b) fluster (c) jumper (d) emergent behavior (e) learning. (p. 4, Dixit & Youngblood, 2008)

Figure 2: Example of Node-link Diagram and Movement Visualization

More recently, Wallner and Kriglstein (2015) discussed the common design strategies useful for comparative data analysis to discover the similarities and dissimilarities of individual differences of learning behaviors in serious games: juxtaposition, superposition, and explicit encoding. These strategies were initially categorized by Gleicher et al. (2011). Using the juxtaposition strategy (see Figure 3), researchers can compare multiple visualizations side-by-side while the superposition strategy overlays the visualizations in a same coordinate system. In contrast, the explicit encoding strategy visualizes any relationships among different data sets such as differences or correlations. Therefore, the explicit encoding helps readers to determine relationships from the visualizations. However, these researchers warned there could be

some challenges for readers such as difficulties with understanding the relationships caused by a lack of prior knowledge of the datasets (Gleicher et al., 2011).

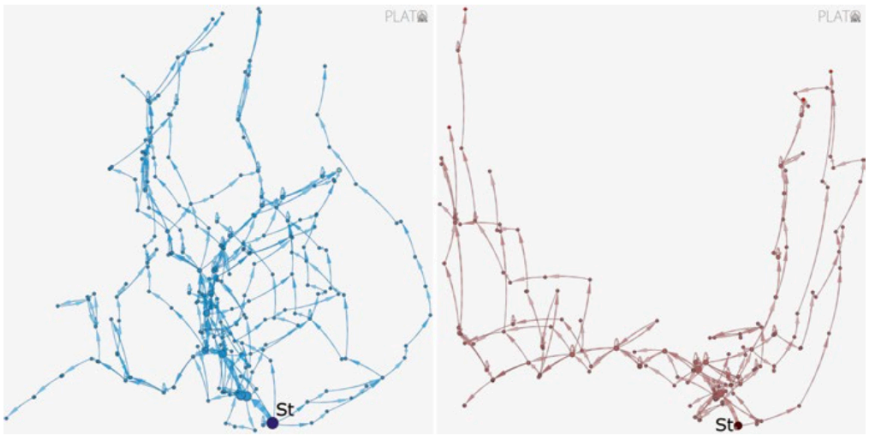


Figure 3: Example of Juxtaposition Strategy (p. 171, Wallner & Kriglstein, 2015)

These various techniques have different purposes and applicability. Kuosa et al. (2016) emphasized that a good visualization can support users to understand data, assure their prior knowledge of a learning environment, and acquire insights for the data. This study intends to explore visualization techniques to illustrate data-driven evidences of students' in-game behaviors, and to develop visualizations that can provide teachers actionable insights by allowing them to track students' actions over time and identify students who may be fallen behind. The findings of this study will support teachers in obtaining an increased practical value that can be applied to facilitate student participation in the context of serious games.

Previous Research on Navigation Patterns in *Alien Rescue*

Research studies have examined students' patterns and usage of cognitive tools in the serious game environment, *Alien Rescue*, through statistical analytics such as

descriptive analysis and cluster analysis. Liu and Bera (2005) investigate the use of cognitive tools across five contextual problem-solving stages (i.e. initial exploring, background research, hypothesis generation, hypothesis testing, and solution generation) through cluster analysis. Results show that students were strategic in their tool usage over each stage. For example, in early stages, students used the tools sharing cognitive process and tools sharing cognitive load more than other tools, which students used more in later stages. Specifically, the students concurrently accessed multiple tools in the later stages. Results indicate that these cognitive tools support students' cognitive skills and facilitate their information processing. The researchers also investigated the relationship between students' use of tools and learning performance and explain that high-performing students tended to use tools in more productive ways than other groups of students. Liu et al. (2009) conducted another study with undergraduate students who played *Alien Rescue* in a laboratory setting. Each student's activities in the environment were observed, and an interview was conducted to determine students' cognitive processes at a specific problem-solving stage. The results support the findings of the previous study—strategic use of cognitive tools throughout the different stages. In comparison to the previous study, this study did not show evidence of different tool usage patterns between high and low performing groups.

Bogard et al. (2013) conducted the descriptive analysis (i.e., cross cluster analysis) using stimulated recall, think-aloud, and direct observation to address how students' application and frequency of cognitive processes and behaviors contributed to differences in performance outcomes and mental model development. The findings revealed that students with consistent self-regulation—the most expert-like learners—kept their cognitive processes on carrying out operations in each threshold of knowledge development: 1) Building a procedural model, 2) building a structural model, 3) building

an executive model, and 4) building arguments. The students developed their mental models through each threshold and thereby focused on the most relevant aspects of the problem solutions. That is, highly self-regulated students tended to evaluate outcomes and readjust their strategies to discover knowledge constraints and build a dynamic mental model of the problem. Therefore, the authors highlighted the role of self-regulation to solve a complex problem and the needs of support for novices' knowledge development.

Liu et al. (2015) applied data visualization techniques to discover sixth graders' usage patterns and identify any contributing factors to student variations. The researchers analyzed students' gameplay data in combination with traditional measures (i.e. self-reported survey of goal-orientation, students' performance score) using the visualization tool *Tableau* (tabluausoftware.com, Computer software, Seattle, WA). *Tableau* specifically represents multidimensional data in a single view. Results indicated different tool usage patterns between different groups of students; for example, high performing and mastery goal-oriented students tended to use the appropriate tools relative to each problem-solving stage.

Overall, existing research confirms the premise that *Alien Rescue* improves students' problem-solving skills and learning and this can be determined via the combination of *in situ* gameplay data with traditional statistical methods (e.g., descriptive analysis, correlation analysis, and cluster analysis) and visualization techniques. Beyond these traditional statistical methods, various data mining techniques hold the promise for discovering meaningful patterns within this open-ended serious game. Further, little is known about the potential meanings of each parameter of *in situ* data as a behavior indicator within this serious game context. This study intends to address these issues.

SUMMARY

This review of literature highlights the relevant theories, issues, methods, and research studies related to serious games environments. Of particular interest is comprehending students' scientific thinking through problem-solving and inquiry as a scientific process within open-ended serious games. Twenty-first century skills, such as critical thinking and problem-solving, are proven to be critical factors for academic or future employment achievements. Advanced technologies with open-endedness in serious game environments facilitate student development of these skills in diverse ways. Attempts to understand how diverse learners (e.g., at-risk and non-at-risk groups or a different level of expertise groups) learn through playing serious games have been made. Serious games analytics increase opportunities for measuring, assessing, and improving students' diverse performance with serious games. Different data mining methods (e.g., k-means cluster analysis, sequential data mining) have been applied to serious games analytics. Data visualization researchers have attempted to understand the differences among individuals using multiple visualization techniques to address interpretive challenges of information derived from the large amount of data.

In reviewing the literature, many issues have been identified that require further investigation. Research stresses that the use of traditional educational assessments is a great challenge in understanding how students learn complex skills through solving scientific problems within open-ended serious game environments. Although open-endedness of a serious game engages students' scientific problem-solving process, their diverse behavior is harder to study because of a game's overall complex system. Therefore, research indicates the importance of using gameplay data (i.e. *in situ* data) and game metrics to better understand individuals' learning behaviors and performances in different contexts. However, extant research is based on traditional assessments such as

pre- and posttests or self-reported surveys in combination with general analytics metrics such as frequency and time-to-completion rate. Particularly within educational contexts, general data mining and visualization techniques must be directed by the theoretical principles about complex learning skills. In addition, insufficient empirical studies have addressed how these techniques can inform pedagogy and inquiry assessment, specifically in an open-ended serious game environment.

A series of studies have examined students' use of cognitive tools in the serious game environment, *Alien Rescue*, and revealed that the cognitive tools within the game support students' cognitive skills and facilitate their information processing. However, we still need to find appropriate metrics that serve to indicate specific skills (such as scientific inquiry) and various techniques of data mining and visualization to assist interpretation of statistical analytics.

Given the challenges of understanding scientific problem-solving processes in open-ended serious games, this study used user-generated data (*in situ* data) in combination with traditional data (*ex situ* data) to understand diverse students' learning behaviors and performances throughout scientific problem-solving in an open-ended serious game environment. In particular, the researcher intends to investigate learning processes among students with at-risk and non-at-risk to identify emergent meaningful patterns by conducting the integrated method of sequential pattern mining and lag sequential analysis. Specific game metrics as an indicator of the scientific inquiry process were used with a data mining technique—*k*-medoids clustering—to identify the patterns and also the relationship between the patterns and learning performance. In addition, diverse visualization techniques such as charts, diagrams, and heatmaps using a juxtaposition strategy were used to support better representation and interpretation of behavior differences to inform teachers just-in-time information of students' in-game

behaviors. The next chapter describes the specific methods employed to further the research in these areas.

Chapter 3: Methodology

Open-ended serious games have the potential to develop students' scientific thinking skills and identify the challenges of understanding students' behaviors because of the complex game systems. Past research has highlighted the importance of using *in situ* gameplay data to examine students' diverse in-game learning behaviors through data mining and visualization techniques. Insufficient empirical studies have addressed how data mining and visualization techniques can be used to investigate students' scientific thinking processes within open-ended serious game environments.

This study seeks to employ statistical methods in combination with data mining and visualization techniques to understand how students solve a problem as they interact with different cognitive tools in an open-ended serious game designed for middle-school science. This study intends, first, to identify learning behavior patterns—as captured by the students' gameplay data—between at-risk and non-at-risk students within the serious game. Then, the study seeks to examine the relationship between students' learning performance and their scientific inquiry behaviors, which emerged as students engaged with Probe Design Center in this serious game. Lastly, this study seeks to explore visualization techniques that can illustrate data-driven evidences of students' in-game behaviors to provide teachers just-in-time support.

This chapter will describe *Alien Rescue*, the research context for this study, the findings of two pilot studies, and the research design employed to address the research questions.

RESEARCH QUESTIONS

The primary purpose of this study is to investigate sixth-grade students' scientific thinking processes in a three-week space science unit with an open-ended serious game environment through using statistical methods in combination with data mining and visualization techniques. This study seeks to investigate the following research questions:

1) Does the average posttest score significantly differ between at-risk and non-at-risk groups?
2) What differences exist between at-risk and non-at-risk students' navigational behaviors as they interact with various in-game tools?
3) What is the relationship between students' scientific inquiry behaviors in Probe Design Center and their learning performance?
4) What scientific inquiry behavior patterns emerge as students engage with Probe Design Center?
5) How can visualizations help to illustrate data-driven evidences of students' in-game behaviors to provide teachers just-in-time support?

PARTICIPANTS

Participants included a convenience sample of 196 sixth graders from a middle school (School A) in the Southwestern area of the United States. The school had used *Alien Rescue* as part of their sixth-grade science curriculum for the past several years. The participants in this study used *Alien Rescue* for six days over three weeks (approximately a total of 500 minutes) on an individual computer; however, the work space encouraged group work. The student demographics were as follows: 13.4% African American, 53.1% Hispanic, 26.1% White, 0.2% Native American, 2.1% Asian, 0.5% Pacific Islander, and 4.7% Two or more races (see Table 2). Gifted and talented students

comprised 6.7% of the populations, 17.3% were enrolled in Special Education, and 15.6% were bilingual/ESL. At-risk students comprised 52.5% of the sample. At-risk students were identified as being at-risk of dropping out of a school based on the state-defined criteria such as low-performance on an assessment instrument and limited English proficiency (Texas Education Agency, 2017).

The teachers provided the students a project checklist presenting each step of problem-solving processes: (1) solar system research, (2) alien species research, (3) elimination chart, (4) probe prototype, (5) probe design, (6) probe launch, (7) probe results, and (8) recommendation. The students filled out a paper-based worksheet for each step that must be approved by the teacher before the students proceeded to next step. The classroom observation revealed more than half of the students spent their time for researching about our solar system during the whole gameplay period; therefore, these students did not make much use of any Probe Design related activities (i.e., Probe Design Center, Mission Control Center). The gameplay data also revealed that only 84 students (42.85%) of School A accessed Probe Design Center. The purpose of the third and fourth research questions is to see the extent to which game metrics generated in Probe Design Center can predict learning performance as being representative of key indicators of students' scientific inquiry behavior in this game context. Due to the lack of sample size and their limited use of Probe Design Center of School A, the researcher included an additional sample of 51 sixth graders from another middle school (School B) in the Southwestern area of the United States to understand students' scientific inquiry behaviors in Probe Design Center across schools and build a model controlling for a school.

School B used *Alien Rescue* as their science curriculum for thirteen days over three weeks (approximately a total of 600 minutes) on an individual computer. The

student demographics were as follows: 3.1% African American, 20.1% Hispanic, 64.2% White, 0.1% Native American, 8.3% Asian, 0.1% Pacific Islander, and 4.1% Two or more races (see Table 2). Gifted and talented students comprised 28.3% of the populations, 12.4% were enrolled in Special Education, and 1.5% were bilingual/ESL.

	School A	School B
Ethnicity		
African American	13.4%	3.1%
Hispanic	53.1%	20.1%
White	26.1%	64.2%
Native American	0.2%	0.1%
Asian	2.1%	8.3%
Pacific Islander	0.5%	0.1%
Two or more races	4.7%	4.1%
Risk Factors		
At-risk	52.5%	28.5%
Economically disadvantaged	60.5 %	7.4%
Limited English proficiency	15.6 %	1.5%
Enrollment by the program		
Bilingual/ESL	15.6%	1.5 %
Gifted and Talented	6.7%	28.3 %
Special Education	17.3%	12.4 %

Table 2: Demographic Information

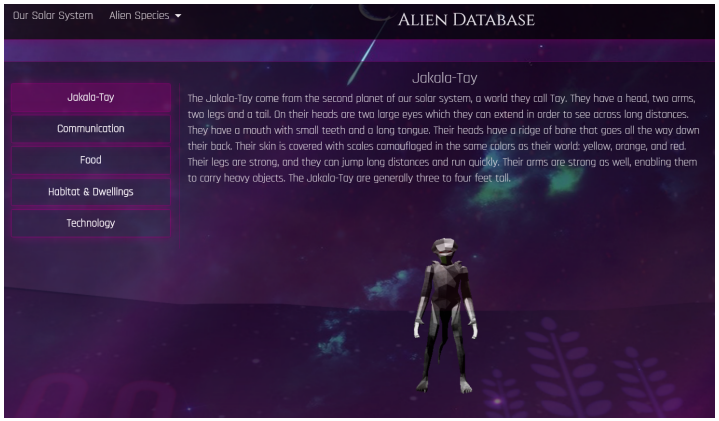
To conduct the study, the approvals from the research site, the parents of the participants, the participants, and the University of Texas at Austin's Institutional Review Board were obtained. The principal of each school confirmed the approval of the research site. Based on the principal's approval, the letter was submitted to the Institutional Review Board. An IRB application was submitted to the Review Board, which consisted of a research proposal, consent letter, assent letter, site approval, and all the surveys to be administered in the study. Approval for all the research participants was obtained, based on the IRB regulations.

RESEARCH CONTEXTS

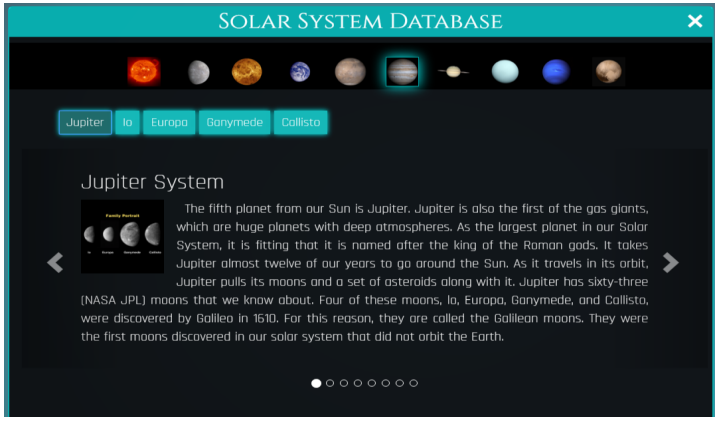
The open-ended serious game, *Alien Rescue* (http://alienrescue.edb.utexas.edu), was developed by a research group consisting of both faculty and graduate students in the Learning Technologies Program at The University of Texas at Austin. Guided by a design-based research framework, this group aspired to generate new theories and improve educational practices using iterative design, development, implementation, and analysis within an authentic real-world setting (Brown, 1992; Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Wang & Hannafin, 2005). During the past decade, *Alien Rescue* has been used as part of the science curriculum by over a dozen middle schools in Central Texas, as well as by schools in at least twenty-nine states and four countries.

Alien Rescue integrates multiple attributes of open-ended serious games along with problem-based learning pedagogy, in which students with different levels of performance use various approaches to solving problems (Glaser, 1991). Authenticity is achieved by placing students in the role of young scientists who are asked to join a United Nations rescue operation to save a group of six distressed aliens displaced from a

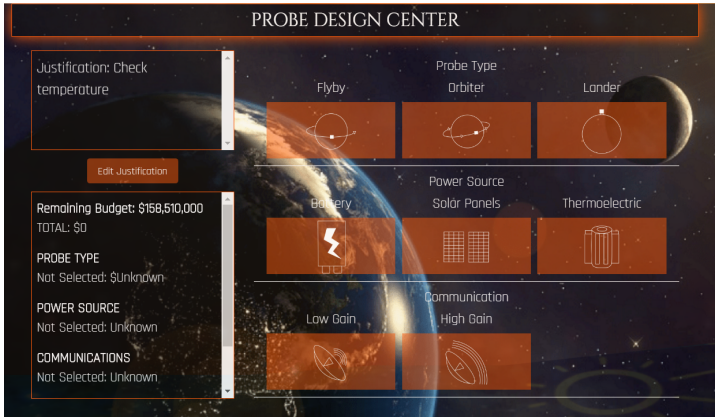
distant galaxy because their home planets have been destroyed. Students are engaged in scientific investigations aimed at the clear goal of finding a suitable home in our solar system in which to relocate each alien species. While students find a solution for each species, they are encouraged to repeat tasks to build a mastery. The central problem of finding the aliens suitable homes is complex, and students are not provided explicit instructions for problem-solving steps. Since this central problem is ill-structured and there are multiple ways to find suitable homes, students need to justify a solution by providing a rationale and evidence. Students explore the multiple functional spaces for supporting cognitive processes, activities, and hypotheses testing, and the affordances of the multiple spaces to develop strategies for utilizing different tools (see Figure 4). Through this open-ended serious game, students experience cognitive processes akin to real-world scientific inquiry and practice high-level cognitive skills such as goal setting, hypothesis generation, problem-solving, and self-regulation.



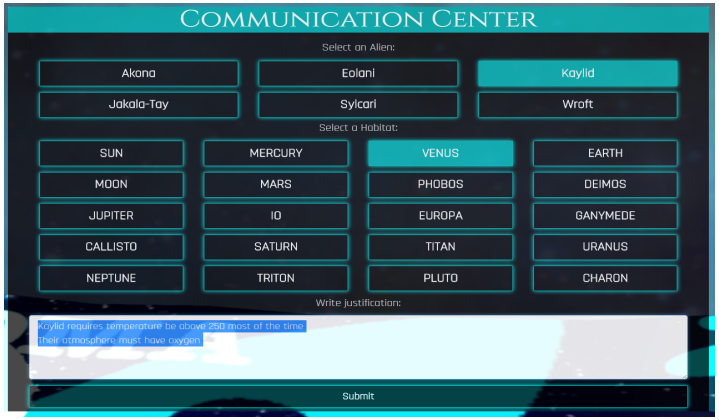
(a) Alien Database



(b) Solar System Database



(c) Probe Design Center



(d) Comunication Center

Figure 4: Screenshots of *Alien Rescue* Environment

Cognitive Tools

To support students' problem-solving processes, *Alien Rescue* provides 10 cognitive tools, each of which have been categorized based on Lajoie's four types of cognitive tool functions (1993; see Table 3): (a) share cognitive load, (b) support cognitive and meta cognitive processes, (c) support cognitive activities that would otherwise be out of reach, and (d) support hypothesis generation and testing. Since each alien has unique needs and characteristics, Students are challenged to gather information embedded in different cognitive tools and integrate this information to solve a complex and ill-structured problem for each alien. Therefore, strategic use of these cognitive tools is essential to complete the students' task. Ten cognitive tools are accessed through a two-layer interface. The first layer consists of four primary tools found in the space station *Paloma*, including Alien Database, Probe Design Center, Mission Control Center, and Communication Center. The second layer consists of the rest of six tools including Solar System Database, Missions Database, Concepts Database, Spectra, Periodic Table, and Notebook, each of which can be overlaid anytime with any tool that students want to access.

Tool categories		Tool functions
Tools sharing cognitive load	Alien Database	Provides descriptions of six aliens' home planets and the characteristics of each species with 3D visuals.
	Solar System Database	Provides (incomplete) information on our solar system that allows students to collect information such as species' habitat.
	Missions Database	Provides information on past NASA missions, including detailed descriptions of probes used on these missions.
	Concepts Database	Provides instructional modules on selected scientific concepts using interactive animations and simulations designed to facilitate conceptual understanding.
	Spectra	Provides information to help students interpret spectra found in the Alien Database.
	Periodic Table	Provides an interactive periodic table of the elements.
Tools supporting cognitive process	Notebook	Allows students to take notes during problem-solving for collecting, summarizing, and integrating information.
Tools supporting otherwise out-of-reach activities	Probe Design Center	Provides an interactive tool for students to design probes that they will send to gather information about planets and moons in our solar system.
Tools supporting hypothesis testing	Mission Control Center	Allows students to review data from launched probes and to integrate information to test hypotheses.
	Communication Center	Provides students with a way to submit their solution for each alien species. Students must also use the form to provide a rationale for their choice of alien habitat. Teachers can review and critique these solutions.

Table 3: Descriptions of Cognitive Tools Provided in *Alien Rescue*

Tools sharing cognitive load

The Alien Database presents the descriptions of the aliens' journey from their home planets to our solar system, and the needs and characteristics of each of the six alien species. This tool provides 3D models of each alien species, their habitats, dietary needs, and technologies. The Solar System Database provides data on selected planets and moons in our solar system; however, this tool is intentionally incomplete and ill-structured, as it does not provide sufficient information to solve the game's problem without further work on the student's part. Therefore, along with the Alien Database, this tool facilitates students' cognitive processes by providing preliminary information with which students generate initial hypotheses and use to iteratively refine their hypotheses as they continue with gameplay. The Missions Database provides information about previous NASA space exploration missions (e.g., Apollo, Galileo space missions). Since this tool provides the purpose, history, and data on scientific instruments of each mission, students use this tool to understand the significance of space missions and how to design a probe. The Concepts Database provides supplemental scientific concepts—atmospheres, temperature, gravity, and supernova—needed during gameplay in an interactive multimedia environment. Therefore, students can use this tool whenever they come across unfamiliar concepts. Lastly, the Spectra tool and Periodic Table support students' interpretations of data found elsewhere in *Paloma*. For example, students can open the Spectra tool to identify specific elements presented in a spectrum of the aliens' habitats.

Tools supporting cognitive process

The Notebook tool supports students' cognitive processes by allowing students a space to organize and compile information from multiple sources while they are working to solve a problem. This tool provides a basic level of scaffolding such as an initial

categorization of note (e.g., a note for aliens, planets, or other information) and specific sub-categories such as atmosphere, temperature, and elements.

Tools supporting otherwise out-of-reach activities

The Probe Design Center helps students to build probes with authentic space exploration technology. This tool particularly provides a novel experience and supports students' authentic scientific inquiry process: to generate and test a hypothesis by building a probe. Students first design a new probe by selecting one or multiple destinations and providing objectives of a mission. Considering the self-identified mission justification, students need to choose a specific probe type among the three options: flyby, orbiter, and orbiter with a lander. Then, students select the power source(s), communication tool(s), and multiple scientific instruments necessary for the probe to gather the desired information. During this process, students are given a limited budget. Therefore, students must be strategic in managing their budget when designing and launching probes. Specifically, certain combinations of instruments, probe types, and/or destinations will produce malfunction errors. It is up to the student to conclude that a specific instrument does not work with a certain probe type or a certain destination. After considering each factor that occurs after a malfunction, the student can revise and retest the hypothesis and probe design. *Alien Rescue* allows students to discover that their choices will impact the data they receive and challenges them to learn from their mistakes and operate strategically. Though the budget limit is not so small that this process is inhibited, if the probe design and launch budget is depleted, funds may be added at the discretion of the teacher.

Tools supporting hypothesis testing

The Mission Control Center allows students to view the data from the chosen launched probe. This tool provides authentic scientific data as graphs or images. Students need to interpret the data to integrate potential information and complete the game's mission. At *Paloma*'s Communication Center, students will receive a message from the Interstellar Relocation Commission Director and submit problem solutions—relocation recommendations for each alien—through the Message Tool.

Findings of Pilot Studies

Previous research examined students' patterns and usage of cognitive tools in the serious game environment, *Alien Rescue*, through statistical analytics and revealed that the cognitive tools within *Alien Rescue* support students' cognitive skills and facilitate their information processing. However, research must continue to seek appropriate metrics as an indicator of a specific skill (e.g., scientific inquiry) and various techniques of data mining and visualization to assist interpretation of statistical analytics.

Research about students' tool usage patterns through data visualization (Liu et al., 2015) used a tool, *Tableau*, with gameplay data in combination with traditional measures (i.e. self-reported survey) to represent multidimensional data in a single view. The study focused on visualizing overall tool usage patterns among different groups by using the average frequency or average total duration of tool use. The findings also brought another attention to an individual's sequence of tool use and sought techniques to visualize the students' learning paths. As a follow-up study, Kang, Liu, and Liu (2017) suggested a visualization technique using D3 (d3js.org), which are more flexible in terms of a variety of ways of graphical representations, but more intensive in terms of technical skill requirements. The researchers first developed a partition algorithm in Python to examine

sixth-grade students' learning paths of six predominant cognitive tools in *Alien Rescue* (i.e. Alien Database, Solar System Database, Probe Design Center, Probe Launch Center, Mission Control Center, and Notebook). Then, the students were divided into seven groups based on their solution scores (e.g., 0-6). In Figure 5, different solution groups are listed inside of the circle. Within each group, a set of bars in an outward direction represents the 50 most frequent tools that students in the group sequentially accessed. Each horizontal bar indicates each tool with its own color. Overall, this visualization showed diverse tool use patterns by different score groups. For instance, the low score groups (i.e., 0-2 score groups) mostly accessed Probe Design and Mission Control, while the high score groups (i.e., 5-6 score groups) used tools relevant to cognitive load and processing at the beginning and then concurrently switched tools. All in all, this findings supported prior research, which indicated different tool use patterns among differently-performing students.

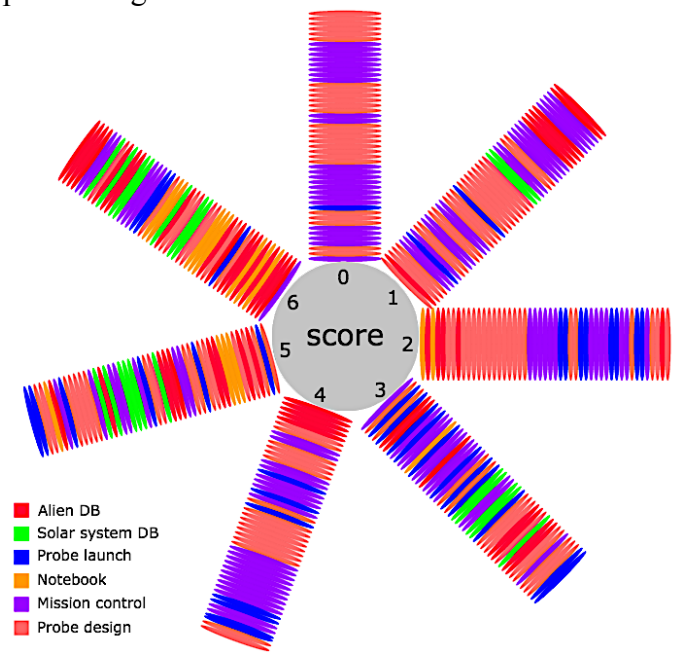


Figure 5: Learning Path of Each Score Group (Kang et al., 2017)

A correlation analysis was conducted to examine the relationship between different metrics from the students' gameplay data (e.g., frequency of each tool use, duration of each tool use) and their learning performance (0-7 score). Results showed two positive behavior indicators of student academic performance: the frequency of Alien Database use, and the time spent on Alien Database. This study confirmed the potential of using two metrics, frequency and duration of tool use, as indicators of learning performance. It also suggested the value of exploring diverse visualization techniques along with multidimensional data since this supports a comprehensive interpretation of the relationships between multiple variables.

Although the frequency and duration metrics showed the potential of the previous studies, other game metrics must be devised as indicators of specific learning behaviors. Another pilot study (Kang & Liu, 2016) was conducted to examine new game metrics, measuring in particular students' scientific inquiry behavior in *Alien Rescue*. This study first examined students' scientific inquiry process patterns in a specific tool, Probe Design Center, which supports students' overall scientific inquiry process by integrating new information and testing a hypothesis via probe design. To measure the inquiry skill, the researchers developed new game metrics: (a) number of launched probes, (b) number of repeated trials, (c) amount of new information, (d) amount of redundant information, and (e) number of errors. The researchers performed cluster analysis to identify scientific inquiry process patterns. The findings of cluster analyses indicated that most of the students—in the bigger cluster—made fewer mistakes toward the later stages. This indicates that the students used a trial-and-error approach during the early stages, but eventually improved their scientific inquiry skills as they approached the final stage.

Ordinal logistic regression was performed for each problem-solving stage to investigate the relationship between learning behaviors and performance. The variables

consisted of the five inquiry skill metrics, the number of a student accessed each tool, and the amount of student usage per each tool as independent variables, and students' submitted solution scores as a dependent variable. The findings revealed a decrease in the numbers of predictors as students approached the final stage. During the early stages, the high performing students tended to devote more time to the tools sharing cognitive load (i.e. Alien Database, Solar System Database). During the later stages, the frequent use of the Probe Design Activities supported their scientific inquiry process of integrating new information and evaluating their own hypotheses. In this study, the researchers developed the logic of each game metric by first observing students' gameplay data and then manually calculating the number of each game metric. To ensure inter-rater reliability, two researchers randomly selected twenty students' gameplay data generated in Probe Design Center and calculated the game metrics of each student to ensure that the same criteria was applied during calculation. During this process, the researchers continuously revised the logic of each game metric until two researchers reached consensus on all logics.

In summary, previous research on *Alien Rescue* has begun to use multiple data sources including *in situ* gameplay data combined with traditional statistical methods (e.g., descriptive analysis, correlation analysis, and cluster analysis) and visualization techniques to understand how students' use of cognitive tools corresponds with different problem solving stages. Although these statistical methods revealed students' behavior patterns in the game, they may not be the optimal methods for discovering meaningful patterns of students' learning processes within this open-ended serious game. Little is known about potential meanings of game metrics driven from *in situ* data as a behavior indicator within this serious game context. Although the pilot study developed new game metrics, the metrics was manually calculated in the study, which highlighted the need of

automated system such as data processing and programming. Thus, built upon the preliminary findings from the pilot studies, this study includes additional techniques to explore new ways of understanding students' learning behaviors and processes.

DATA SOURCES

In this study, the researcher investigated students' problem-solving and inquiry as a scientific thinking process within an open-ended serious game. To address this, the study intends to employ both statistical analytics and unsupervised learning techniques (sequential pattern mining, k-means cluster analysis) by using *in situ* data (i.e. user-generated data from the cognitive tools, solution texts) in combination with *ex situ* data (i.e. science knowledge test). The following sections outline each data source.

In Situ User-Generated Data

Navigation Data

The overall navigation data—that is, the user-generated data from all cognitive tools—was used to identify students' behavior patterns as they engaged in *Alien Rescue*. The game records every action as each student interacts with the environment. Data contains a student identifier, a cognitive tool that a student accesses, a type of action a student is taking (e.g., open or close), any additional notes on the student's interactions, and a timestamp for each action (see an example in Table 4).

This raw gameplay data makes it possible to determine students' activities in-game such as sequence, frequency, and duration of cognitive tool use. Each tool generates different actions based on its function. For example, students can zoom in and out of 3D models of each alien species in Alien Database, which is not an available

feature in other cognitive tools. Therefore, to determine each student's sequential cognitive tool use, the researcher only included an open action, which was required to determine a sequence and frequency of in-game tool use. For instance, the sequence of cognitive tool use for the student with the 147893 identifier in Table 4 is {Probe Design Center, Probe Design Center, Alien Database}. The frequency of Probe Design Center is 2 based on the student's open actions.

Student ID	Tool	Action	Timestamp	Notes
147893	probe design	open	5/24/16 14:34	
147893	probe design	close	5/24/16 14:35	
147893	probe design	open	5/24/16 14:36	
147893	probe design	open probe	5/24/16 14:36	
147893	probe design	close	5/24/16 14:39	
147893	alien database	open	5/24/16 14:35	
147893	alien database	close	5/24/16 14:35	
147893	spectra	open	5/24/16 14:35	
147893	periodic	open	5/24/16 14:35	
147893	spectra	close	5/24/16 14:35	
147893	spectra	open	5/24/16 14:35	
147893	alien database	open	5/24/16 14:35	
147893	solar system	open	5/24/16 14:35	
147893	solar system	click	5/24/16 14:36	Jupiter System
147893	solar system	click	5/24/16 14:36	Io
147893	solar system	click	5/24/16 14:38	Europa

Table 4: Example of Navigation Data

Data was cleaned by including only meaningful gameplay data. First, the researcher observed a specific period for each class. For example, the gameplay data not recorded during the classroom sessions were removed, since some students accessed the program after school or during the holidays. The researcher used Python to transform the navigation data into sequence data to perform sequence analyses. The navigation data were then translated into a vertical id-list database as an input file including a student ID, event ID, item size and item/tool(s) (see Table 5). An event ID indicates a sequential order of each action of a student, and an item size indicates the number of tools used in each event.

Student ID	Event ID	Item size	Tool name
147893	1	1	alien database
147893	2	1	probe design
147893	3	1	notebook
147893	4	1	alien database
147893	5	1	solar database
147893	6	1	solar database
147893	7	1	spectra
147893	8	1	periodic table
147893	9	1	probe design
147893	10	1	alien database
147893	11	1	alien database
147893	12	1	solar database

Table 5: An Example of Input File for Sequence Analyses.

Probe Design Activity Data

One of the cognitive tools in *Alien Rescue* is Probe Design Center. This interactive tool provides students a novel experience, which allows the students to design probes using authentic space exploration instruments that will return important information about planets and moons in our solar system. Probe Design Center supports students' scientific inquiry process to generate and refine their own hypotheses. Students first need to select a destination(s) out of the 19 planets in our solar system and justify a hypothesis. Then, they choose one of the following probe types: flyby, orbiter, or orbiter with a lander. Finally, they select a power source, communication tool, and scientific instruments to install on their probe.

During this process, students are given a limited budget; therefore, they need to consider how to manage their budget. They might decide not to install a scientific instrument that would return information already provided in the Solar System Database, or that would cause malfunction in a certain world. For example, installing a seismograph—an instrument to detect any seismic activity—on a probe headed for any gas giant like Jupiter would cause an error, since a seismograph only works on worlds with hard surfaces. Therefore, this tool encourages students to refine their design process. Students' design decisions directly influence data available later at the Mission Control Center. However, research on scientific problem-solving indicates that young students have difficulties in conducting scientific inquiry (Gobert et al., 2015; Hmelo-Siver & Azevedo, 2006; Lederman et al., 2014). For example, when designing experiments, students often collect limited evidence to test a hypothesis, attempt a few trials or repeated trials without changing any variables or conditions, or change too many variables (Gobert et al., 2015). Therefore, game metrics that can measure students' specific behaviors need to be developed to better understand how diverse students

conduct scientific inquiry and facilitate their inquiry process in a complex learning environment such as *Alien Rescue*.

Students' gameplay data in Probe Design Center contains every instance of action when a student builds a probe. Data includes a user identifier, probe name, selected destination, written hypothesis, probe launched (or not), selected probe type (i.e. flyby, orbiter, orbiter with lander), selected instruments, and timestamp. To understand students' different inquiry processes and strategies using the Probe Design Center, five game metrics of measuring scientific inquiry skills were defined: (a) *number of launched probe*s, (b) *number of repeated trials*, (c) *amount of new information*, (d) *amount of redundant information*, and (e) *number of errors* (see Table 6).

Metrics	
Number of launched probes	Number of launched probes
Number of repeated trials	Number of launched probes to the same destinations with any previous probes
Amount of new information	Number of instruments that return new data
Amount of redundant information	Number of instruments that return information that can be found elsewhere (e.g., Solar System Database)
Number of errors	Number of instruments that return errors

Table 6: Game Metrics of Measuring Scientific Inquiry Skills in Probe Design Center

First, the *number of launched probes* metric simply counts the number of probes launched by each student. Second, if a student launched a probe to the same destination(s) in addition to previous probes, this action counts as a *repeated trial*. Third, the *amount of new information* metric counts a number of instruments that return new data with a lander. Fourth, the *amount of redundant information* metric counts the

number of instruments that return information obtainable elsewhere (e.g., Solar System Database) or the number of instruments already selected in the previous trial. Lastly, the *number of errors* metric counts instruments that return errors. For example, spectrograph, seismograph, thermometer, and barometer in any flyby or orbiter probe cause an error, since these instruments only work properly with a lander with orbiter probe. As previously explained, a seismograph on a lander probe to any gas giant (i.e. Jupiter, Saturn, Uranus, Neptune) will cause an error. Accordingly, all instruments only with the probe type, a lander, can gather the data without any malfunction; therefore, only a lander probe type was only considered when counting the amount of new information. In this analysis, only launched probes were considered because students can continue revising a probe before launching it. Additionally, any probes launched to Earth or the Sun will be removed, since the probes will malfunction. Based on these definitions, a Python script was developed to calculate each metric. Appendix A contains the matrix of the amount of new information, amount of redundant information, and the number of error variables.

Problem Solutions

Students' solutions were evaluated by how successfully a student solves the central problem. Students use the Message Tool to submit a solution(s) for each alien, and they must indicate an appropriate home for each species and provide a rationale (see the example in Table 7). Students can submit multiple solutions for each alien species, which reveals the results of students' problem-solving processes—that is, justifications of their solutions using the gathered data. In a real classroom environment, students move through the problem-solving processes at their own pace. Therefore, some students can submit all solutions for six alien species while others cannot. In addition, since there are

multiple possible answers and multiple suitable homes for each alien, students can submit multiple solutions for each alien species.

The solutions were evaluated using an 8-point rubric (Appendix B) used in previous studies (Bogard, Liu, & Chiang, 2013; Liu et al., 2009) in terms of the correctness of the solution and the number of reasons to the selected home. Two key criteria determine scores between 1-8: first, the feasibility of worlds students select (certain planets or moons are inhabitable while other worlds are uninhabitable choices given the characteristics of the alien species and the planets), and second, the number of reasons students provide to prove their choice of world. Students who recommend an uninhabitable planet are given a score of 1. A score of 2 is given to students who recommend an inhabitable world but provide one reason to explain their choice. An additional point is granted for each reason students provide that is informed by the data analysis they conduct while working in *Alien Rescue*. For example, students who provide two reasons for their choice get 3 points and students who provide three reasons get 4 points. The maximum score of 8 is granted to students whose solutions provide six or more reasons and provide constraints of the proposed home or address how the limitations can be controlled.

Two researchers went through the entire scoring process to get 100% inter-rater reliability. The researchers worked together at first to score 10% of solutions to ensure that they applied the same criteria for scoring. Then, each researcher scored the remainder of solutions independently.

Student ID	Alien	Destination	Justification	Timestamp
124959	Eolani	Venus	The temperature matches they can live on Earth and Venus is almost identical to Earth and Venus has Oxygen and it's close to the sun and good seismic activity and the Atmosphere is Oxygen so they can live on it.	5/27/16 10:31
5294	Jakala-Tay	Io	The atmosphere is good enough for them to live in the tempertautre is pretty nice and the magnetic feild is strong and they can survive the seismic activity and craters.	6/01/16 10:35

Table 7: Example of Students' Problem Solutions

Ex Situ Data

Space Science Knowledge Test (SSKT)

Learning performance was measured by students' comprehension of the various scientific concepts introduced in *Alien Rescue*. Twenty-four multiple-choice questions (Cronbach's alpha = 0.768) were administered before and after gameplay. This science knowledge test addresses both factual and applied knowledge in the game. The total range of the scores for this test is between 24 and 0. Each question has four answer choices, including a "not sure" option. Examples of each types of questions (i.e. factual and applied question) are as following:

Factual: Which of these worlds is a gas giant?

A. Saturn

B. Earth

C. Pluto

D. Not sure

Applied: Suppose that you want to take close-up pictures of features on the surface of Callisto, but you can only afford to send an orbiter. What instrument would you include?

A. Infrared camera

B. Narrow angle camera

C. Barometer

D. Not sure

ANALYSIS

Identifying Navigation Behavior Patterns

This study was particularly interested in understanding students' navigation behavior patterns during their problem-solving process. Extant studies on cognitive processes in *Alien Rescue* used traditional statistical analysis including descriptive analysis. This study expanded on previous research by incorporating statistical and data mining techniques in the investigation of learning behaviors within a game environment. Two research questions were asked to identify navigation behavior patterns:

1) Does the average posttest score significantly differ between at-risk and non-at-risk groups?
2) What differences exist between at-risk and non-at-risk students' navigational behaviors as they interact with various in-game tools?

To answer the first research question, a One-Way ANCOVA was conducted on a dependent variable: SSKT posttest score. The independent variable was the at-risk classification (at-risk and non-at-risk). The SSKT pretest score was used as the covariate. As for the second research question, a Mann-Whitney U test was performed to determine if there were differences in the frequencies of each in-game tool use between at-risk and non-at-risk groups.

Previous research revealed the problem-solving process used in this game can be grouped into different stages (e.g., Kang, Liu, & Qu, 2017; Liu et al., 2015, 2016). However, the classroom observation revealed the school in this study used the game only six days, and the students played the game longer in each day, while the other schools in the previous studies used the game approximately 10-15 days. Therefore, this study considered each day as one single stage of problem-solving process. The researcher used the navigation data to count each student's daily frequency of each tool use. Then, the frequencies of each tool use by each day (a total of six days) were treated as an individual variable; that is, a total of 6 variables for each in-game tools. Then, Lag Sequential Analysis (LSA; Bakeman & Gottman, 1997) and a sequential pattern mining (Agrawal & Srikant, 1995) were conducted to further discover navigational behavior patterns between the non-at-risk and at-risk students to identify the most appropriate way of profiling students' behavior patterns in this game context. The researcher performed both cSPADE and LSA using the vertical format of navigation data (see Table 5) by each group (i.e., non-at-risk and at-risk) each day.

First, LSA on each group's chronological tool use data (see Table 5) of each day was conducted to obtain daily sequential behavior patterns for each group. LSA yields a series of sequential analysis matrix calculations: transitional frequency matrix, transitional probability matrix, and adjusted residuals table (Bakeman & Gottman, 1997).

In this study, a frequency of transition indicates the number of occurrences of transition from one tool to the consecutive tool among each group's tool use data of each day. As shown in Table 8, considering only three of the ten in-game tools in *Alien Rescue*, Alien Database is followed by each of these tools: Alien Database 15 times, Solar System Database 20 times and Missions Database 5 times. A transitional probability indicates the likelihood that an initial tool follows a subsequent or same tool; that is, the occurrence for each cell (i.e., each transition) divided by the occurrence for that row. For example, the transitional probability of Solar System Database, given that AlieTrann Database just occurred, is 20/40 = .50, indicating that Solar System Database followed Alien Database 50% of the time in the sequences of the group of students.

		Target tool			
		Alien DB	Solar DB	Missions DB	Totals
Given tool	Alien DB	15	20	5	40
	Solar DB	10	25	5	40
	Missions DB	0	5	10	20
	Totals	25	50	20	100

Table 8: Example of Observed Frequencies for Two-item Sequences.

Adjusted residuals (i.e., z-scores) of each transition were calculated to determine if the transitional probabilities deviated significantly from the expected value. The adjusted residual was calculated using the following formula (Bakeman & Gottman, 1997):

$$z_{GT} = \frac{o_{GT} - e_{GT}}{\sqrt{e_{GT}(1 - {o_{G+}}/{N})(1 - {o_{+T}}/{N})}}$$

where o_{GT} is an observed value for the transition from given (G) to target tool (T), e_{GT} is an estimate of the expected frequency, $o_{G+}$is the total observed frequencies in the G-th row, o_{+T} is the total observed frequencies of the T-th column, and N is the total number of frequencies in the table. An adjusted residual identifies any transitions that occur at frequencies greater than chance; for example, a z-score above 2.32 denotes the behavioral transition from a tool to another in the process of problem-solving in a group of students reaches a significant level of 0.01 ($p < .01$).

Next, cSPADE algorithm (Zaki, 2000, 2001) was applied to perform a sequential pattern mining. The cSPADE algorithm discovers constrained frequent sequences; that is, in this present study, frequent sequences of tools among students. The algorithm uses a vertical id-list database as an input file where each transaction includes an object ID, event ID, item size and item(s). The researcher specified two user-specified thresholds: a minimum support of .25 (*minsup*; indicating the results only show the sequences that more than 25% of students used) and a maximum gap of 2 (*maxgap*; specifying the maximum time difference between consecutive elements of a sequence). To compare cSPADE with LSA, the researcher specified 2 as *maxgap* (i.e., time difference), since LSA only looks up two consecutive items in this study (i.e., *lag* = 1; Bakeman & Gottman, 1997). *lag* 1 indicates the transitions from one action to the subsequent action (Pohl et al., 2015). This study examined only significant two-action sequences (i.e., *lag* 1) since the previous studies found students do not typically make many actions in this game context because of the limited amount of time spent in using the game (Kang et al., 2017; Liu et al., 2015, 2016).

Effect on Science Knowledge

This study is further interested in understanding students' inquiry behavior patterns of using a tool, Probe Design Center, which supports students' scientific inquiry process by generating and refining hypotheses and designing their probes. Two research questions were asked to examine the effect of students' inquiry behavior patterns on science knowledge:

3) What is the relationship between students' scientific inquiry behaviors in Probe Design Center and their learning performance?
4) What scientific inquiry behavior patterns emerge as students engage with Probe Design Center?

In order to investigate students' scientific inquiry behaviors in Probe Design Center, five game metrics for measuring scientific inquiry skills were used: (a) number of launched probes, (b) number of repeated trials, (c) amount of new information, (d) amount of redundant information, and (e) number of errors (see Table 6). As mentioned in the Participants section above, the classroom observation and the gameplay data revealed that only 84 students (42.85%) accessed Probe Design Center, since the students were not allowed to proceed to use Probe Design Center until they submitted their paper worksheets of the research on our solar system, in which the students needed to fill in at least five factors of each planet and the moons. Therefore, many students spent most of their time for researching about our solar system and eventually could not make much use of any Probe Design related tools (i.e., Probe Design Center, Mission Control Center). Compared with the other schools previously observed (Kang et al., 2017; Liu et al., 2015, 2016), the students in this present study showed less access to Probe Design Center.

To discover game metrics as a key indicator of learning in this game context, which can be representative across schools, the researcher included additional data from School B (see Table 2) to build a model controlling for a school. In addition, since School B used the game longer than School A, the researcher considered the number of launched probes variable as another covariate, which could confound the regression results. That is, the researcher included the four predictors: (1) number of errors, (2) number of repeated trials, (3) amount of new information, and (4) amount of redundant information, and the two covariates: (1) school and (2) number of launched probes. To investigate both in-game and after-game learning performances, the researcher conducted two hierarchical regression analyses on: (1) average solution scores (a total solution score / a total number of solution submissions) as in-game performance, (2) SSKT posttest scores, as a dependent variable for each analysis. The participants of School B only took the SSKT posttest, not the pretest; therefore, the researcher only considered the posttest scores as after-game learning performance.

To address the fourth research question, a cluster analysis was carried out to discover inquiry behavior groups in each school. Five game metrics were entered to conduct an unsupervised classification analysis (i.e. *k*-medoids cluster analysis) to identify the potential groups of students with similar behavior traits. This study conducted a *k*-medoids clustering using the partitioning around medoids (*pam*) algorithm in *R* (Kaufman & Rousseeuw, 1990). Both k-means and *k*-medoids are the partitioning algorithms, which attempt to minimize the distance between observations assigned to be in a cluster. However, in *k*-means a center of each cluster is the average of observations in the cluster, while in *k*-medoids each center is one of the observations itself. That is, a medoid can be defined as the observation of a cluster whose average dissimilarity to all the observations in the cluster is minimal. Similar with the *k*-means method, *k*-medoids

also requires a predefined number of clusters. With the *k*-means method the centers of the clusters are only recalculated after all the observations have moved from one cluster to another. On the other hand, *k*-medoids constantly recalculate the sums of the distances between objects within a cluster as observations move around, which requires considerably more computation than the *k*-means method, but leads possibly a more reliable solution. The *k*-medoids method is more robust to outliers and noise as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances. Additionally, to examine the fuzziness of the data, this study conducted a fuzzy clustering, which is appropriate to analyze log data when prior information is little or unknown (Jain, Murty, & Flynn, 1999). The fuzziness of the data was examined by the normalized Dunn coefficients (ranging from 0 to 1): 0.949 (School A), 0.965 (School B), indicating the data is not fuzzy. That is, a fuzzy cluster analysis is not necessary, and the data can be well-partitioned with a hard cluster analysis such as *k*-medoids (Kerr & Chung, 2012).

Since a cluster analysis can only reveal the latent cluster patterns of learning behavior, a Kruskal-Wallis H test was applied to ascertain a statistical significance of the behavioral differences between clusters. In addition, the cluster patterns were visualized for deeper understanding of learning behaviors in each cluster.

Visualization for Just-in-time Support

This study was particularly interested if visualizations can be used to support teachers to provide the students just-in-time feedback. The researcher conducted the classroom observation to monitor how students engage in using diverse in-game tools in *Alien Rescue* and how teachers interact with students and facilitate students' problem-

solving processes. The following research question on visualization was asked to address the challenges of understanding students' problem-solving processes through information interpretation derived from massive user-generated data in this game:

5) How can visualizations help to illustrate data-driven evidences of students' in-game behaviors to provide teachers just-in-time support?

Data visualization techniques can address the challenges to information interpretations derived from large amounts of user-generated data. Diverse graphical representation techniques support better comprehension of differences among student groups. Different visualization techniques embedded in *Tableau* such as charts, diagrams, and heatmaps using a juxtaposition strategy were applied to deliver insight about diverse students' in-game behaviors. The visualizations were proposed as an example of teachers' interactive dashboard to help teachers to track students' participation and identify students who have fallen behind, which ultimately enables teachers to facilitate students' cognitive processes in the context of this serious game.

Chapter 4: Results

IDENTIFYING NAVIGATION BEHAVIOR PATTERNS

The purpose of this study is to identify students' navigation behaviors in cognitive processes in *Alien Rescue*. This study incorporated statistical and data mining techniques using gameplay data to investigate in-depth navigation behavior patterns between at-risk and non-at-risk students within this game environment. Two research questions were asked to identify navigation behavior patterns:

1) Does the average posttest score significantly differ between at-risk and non-at-risk groups?
2) What differences exist between at-risk and non-at-risk students' navigational behaviors as they interact with various in-game tools?

Research Question 1: Does the average posttest score significantly differ between at-risk and non-at-risk groups?

Little is known about at-risk students' cognitive processes or challenges in open-ended serious games such as *Alien Rescue* that require higher-order thinking skills such as reasoning through problem-solving. As mentioned in the participants section, the classroom observations revealed the challenges of at-risk students' cognitive processes in Alien Rescue, in which most students spent their time for researching about our solar system using the information in Solar System Database during the entire gameplay period. Therefore, the first research question addressed what differences exist between at-risk and non-at-risk students' navigation behavior patterns.

The researcher first investigated whether one group has a higher SSKT posttest score mean after the gameplay; that is, the differences in the SSKT posttest means after accounting for the pretest scores between at-risk and non-at-risk groups. One-Way ANCOVA was conducted with an independent variable of at-risk classification (at-risk vs. non-at-risk) and a dependent variable of SSKT posttest score. The SSKT pretest score was used as the covariate. Before conducting the ANCOVA to verify the effects of the variables, requirements and assumptions were confirmed. First, the independent variable was categorical variable. Second, no significant outliers were found. Third, examining a Shapiro-Wilks test on the residuals indicated the residuals were approximately normally distributed for each category of the independent variable. Fourth, a diagnostic test to assess the homogeneity of regression slopes was performed. The test evaluates the interaction between the covariate (i.e., pretest score) and the independent variable (i.e., at-risk classification) in the prediction of the dependent variable (i.e., posttest score). A significant interaction between the covariate and the factor suggests that the differences on the dependent variable among groups vary as a function of the covariate. The interaction source is labeled At-risk*Pretest (see Table 9). The results suggested the interaction was not significant, $F(1, 192) = .425$, $p = .515$ (i.e., $p > 0.05$). Fifth, there was homogeneity of variances, as assessed by $F(1, 194) = .579$, $p = .447$ (i.e., $p > 0.05$). Overall, there was no significant violation of assumption to implement ANCOVA, therefore, the researcher proceeded with the ANCOVA analysis.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2418.743[a]	3	806.248	89.830	.000	.584
Intercept	676.799	1	676.799	75.407	.000	.282
At-risk	9.156	1	9.156	1.020	.314	.005
Pretest	1300.670	1	1300.670	144.917	.000	.430
At-risk * Pretest	3.816	1	3.816	.425	.515	.002
Error	1723.257	192	8.975			
Total	34418.000	196				

[a]R Squared = .584 (Adjusted R Squared = .577)

Table 9: Tests of Between-Subjects Effects on SSKT Posttest Score

The One-Way ANCOVA results revealed there was a significant difference in mean SSKT posttest score ($F(1, 193) = 16.911$, $p < .01$, partial eta-squared (η_p^2) = .081) between non-at-risk and at-risk groups (see Table 10). Table 11 shows the adjusted SSKT posttest score means controlling for the covariate SSKT pretest score for each group. Comparing the estimated marginal means showed that the non-at-risk group showed significantly different scores on the SSKT posttest score ($M = 13.34$, $SE = .31$) compared to the at-risk group ($M = 11.40$, $SE = .33$). That is, the non-at-risk group showed significantly higher improvement on science knowledge than the other group. However, the effect size of at-risk classification is small (partial eta squared (η_p^2) = .081) by the rule of thumb (Cohen, 2013; MRC Cognition and Brain Sciences Unit, 2009).

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2414.926[a]	2	1207.463	134.934	.000	.583
Intercept	673.474	1	673.474	75.261	.000	.281
Pretest	1455.384	1	1455.384	162.639	.000	.457
At-risk	151.331	1	151.331	16.911	.000	.081
Error	1727.074	193	8.949			
Total	34418.000	196				
Corrected Total	4142.000	195				

[a]R Squared = .583 (Adjusted R Squared = .579)

Table 10: Tests of Between-Subjects Effects on SSKT Posttest Score without Interaction Term

Independent Variable	Dependent Variable (SSKT Posttest Score)		95% Confidence Interval	
	Mean	Std. Error	Lower Bound	Upper Bound
Non-at-risk	13.337	.307	12.731	13.944
At-risk	11.401	.329	10.753	12.050

Table 11: SSKT Estimates by At-Risk Classification

Research Question 2: What differences exist between at-risk and non-at-risk students' navigational behaviors as they interact with various in-game tools?

Considering the previous result that the non-at-risk group had a higher SSKT posttest mean score after the gameplay, the researcher further examined what behavior

patterns emerged as students in at-risk and non-at-risk group interacted with various in-game tools during their problem-solving process.

Daily Frequencies of in-game tool uses

The researcher first addressed whether the daily frequencies of each in-game tool use differed between at-risk and non-at-risk students. The variables were evaluated for the normal distribution. The Shaprio-Wilk statistic showed the data were not normally distributed for all variables ($p < 0.05$). Since ANOVA is robust against deviations from normality, the researcher evaluated the violation of assumptions for non-parametric test. First, all dependent variables were measured at the continuous level. Second, the independent variable consisted of two categorical, independent groups: at-risk and non-at-risk groups. Third, the data set met the independence of observations assumption; that is, there was no relationship between the observations in each group or between the groups themselves. Lastly, two distributions of each tool use frequencies for both groups (i.e., at-risk and non-at-risk groups) showed the same or at least similar shape. Overall, there was no significant violation of assumptions to implement a non-parametric statistical analysis, Mann-Whitney U test.

A Mann-Whitney U test was performed to determine if there were differences in the frequencies of each in-game tool use between at-risk and non-at-risk groups. Table 12 shows that the results of the Mann Whitney U-test applied to the frequencies of each in-game tool use of the students in the non-at-risk and at-risk groups. Bonferroni correction (Bland & Altman, 1995) was used to control for the inflated family-wise Type I error rate when counteracting the problem of multiple comparisons (e.g., n hypotheses; a number of gameplay days for each tool) at a statistical significance level of α/n (i.e., 0.05/6 = 0.0083 in this study). The results revealed a statistically significant differences of Alien

Database use ($p < .0083$) on Days 4-6 ($U_{Day4} = 3707.500$, $Z_{Day4} = -2.907$; $U_{Day5} = 3786.000$, $Z_{Day5} = -2.877$; $U_{Day6} = 3811.500$, $Z_{Day6} = -2.698$). During these days, the rank averages of the tool use frequencies of the non-at-risk group students were higher than the students in the at-risk group. The results also indicated the frequencies of Communication Center use during Days 5 and 6 of the non-at-risk students were significantly higher that the students in the at-risk group ($U_{Day5} = 3814.000$, $Z_{Day5} = -2.900$; $U_{Day6} = 3809.000$, $Z_{Day6} = -2.712$). Similarly, the non-at-risk group showed significantly high use of Probe Design Center than the other group during Days 4-6 ($U_{Day4} = 3818.000$, $Z_{Day4} = -2.665$; $U_{Day5} = 3398.500$, $Z_{Day5} = -3.782$; $U_{Day6} = 3108.500$, $Z_{Day6} = -4.345$). As for the frequencies of Mission Control Center use, the results revealed statistically significant differences between the non-at-risk and at-risk groups during Days 5-6 ($U_{Day5} = 3536.000$, $Z_{Day5} = -3.956$; $U_{Day6} = 3461.000$, $Z_{Day6} = -3.748$). The non-at-risk group also showed significantly high use of Spectra than the other group on Day 4 ($U_{Day4} = 3736.500$, $Z_{Day4} = -3.345$).

The analyses showed no significant difference between the rank averages of the groups' frequencies of Solar System Database use during the most of the days except for Day 4 ($U_{Day4} = 3785.000$, $Z_{Day4} = -2.641$, $p = 0.008$). The rank average of the frequencies of tool use of the non-at-risk group students was 108.11, while the students in the at-risk group had a frequency rank average of 87.64. Similarly, the non-at-risk group students used Spectra significantly more only on Day 4 ($U_{Day4} = 3736.500$, $Z_{Day4} = -3.345$, $p = 0.001$). The rank average of the frequencies of Spectra use of the non-at-risk group students was 108.57, while the students in the at-risk group had a frequency rank average of 87.11.

Interestingly, the analyses had shown almost no significant difference between the rank averages of the groups' frequencies of tool use on the early days; however, an

examination of the rank averages of their frequencies of tool use during the later days demonstrated that the students in the non-at-risk group tended to interact with all in-game tools more often than those in the at-risk tools. This result indicates that problem-solving strategies were differently used between the non-at-risk and at-risk students within this environment.

Tools	Days	Groups	Mean Rank	Sum of Ranks	U	Z	*p*
Alien Database	1	Non-at-risk[a]	101.88	10595.50	4432.500	-.927	.354
		At-risk[b]	94.68	8710.50			
	2	Non-at-risk	104.44	10862.00	4166.000	-2.056	.040
		At-risk	91.78	8444.00			
	3	Non-at-risk	99.38	10335.00	4693.000	-.265	.791
		At-risk	97.51	8971.00			
	4	Non-at-risk	108.85	11320.50	3707.500	-2.907	.004*
		At-risk	86.80	7985.50			
	5	Non-at-risk	108.10	11242.00	3786.000	-2.877	.004*
		At-risk	87.65	8064.00			
	6	Non-at-risk	107.85	11216.50	3811.500	-2.698	.007*
		At-risk	87.93	8089.50			
Communication	1	Non-at-risk	95.74	9956.50	4496.500	-.770	.441
		At-risk	101.63	9349.50			
	2	Non-at-risk	95.41	9923.00	4463.000	-1.037	.300
		At-risk	101.99	9383.00			

Table 12 (continued)

	3	Non-at-risk	98.68	10262.50	4765.500	-.066	.948
		At-risk	98.30	9043.50			
	4	Non-at-risk	106.25	11049.50	3978.500	-2.487	.013
		At-risk	89.74	8256.50			
	5	Non-at-risk	107.83	11214.00	3814.000	-2.900	.004*
		At-risk	87.96	8092.00			
	6	Non-at-risk	107.88	11219.00	3809.000	-2.712	.007*
		At-risk	87.90	8087.00			
Mission Control Center	1	Non-at-risk	97.24	10112.50	4652.500	-.388	.698
		At-risk	99.93	9193.50			
	2	Non-at-risk	95.49	9930.50	4470.500	-1.322	.186
		At-risk	101.91	9375.50			
	3	Non-at-risk	93.16	9689.00	4229.000	-2.151	.032
		At-risk	104.53	9617.00			
	4	Non-at-risk	105.00	10920.50	4107.500	-2.230	.026
		At-risk	91.15	8385.50			
	5	Non-at-risk	110.50	11492.00	3536.000	-3.956	.000*
		At-risk	84.93	7814.00			
	6	Non-at-risk	111.22	11567.00	3461.000	-3.748	.000*
		At-risk	84.12	7739.00			
Probe Design Center	1	Non-at-risk	93.41	9714.50	4254.500	-1.767	0.077
		At-risk	104.26	9591.50			
	2	Non-at-risk	101.04	10508.00	4520.000	-1.193	0.233
		At-risk	95.63	8798.00			

Table 12 (continued)

	3	Non-at-risk	101.25	10529.50	4498.500	-0.988	0.323
		At-risk	95.40	8776.50			
	4	Non-at-risk	107.79	11210.00	3818.000	-2.665	0.008*
		At-risk	88.00	8096.00			
	5	Non-at-risk	111.82	11629.50	3398.500	-3.782	0.000*
		At-risk	83.44	7676.50			
	6	Non-at-risk	114.61	11919.50	3108.500	-4.345	0.000*
		At-risk	80.29	7386.50			
Solar System Database	1	Non-at-risk	95.79	9962.50	4502.500	-0.753	0.451
		At-risk	101.56	9343.50			
	2	Non-at-risk	98.91	10287.00	4741.000	-0.114	0.910
		At-risk	98.03	9019.00			
	3	Non-at-risk	99.10	10306.00	4722.000	-0.167	0.868
		At-risk	97.83	9000.00			
	4	Non-at-risk	108.11	11243.00	3785.000	-2.641	0.008*
		At-risk	87.64	8063.00			
	5	Non-at-risk	95.99	9983.00	4523.000	-0.734	0.463
		At-risk	101.34	9323.00			
	6	Non-at-risk	99.18	10314.50	4713.500	-0.203	0.840
		At-risk	97.73	8991.50			
Spectra	1	Non-at-risk	95.89	9972.50	4512.500	-1.205	0.228
		At-risk	101.45	9333.50			
	2	Non-at-risk	99.59	10357.00	4671.000	-0.661	0.508
		At-risk	97.27	8949.00			

Table 12 (continued)

3	Non-at-risk	100.02	10402.50	4625.500	-0.668	0.504
	At-risk	96.78	8903.50			
4	Non-at-risk	108.57	11291.50	3736.500	-3.345	0.001*
	At-risk	87.11	8014.50			
5	Non-at-risk	102.76	10687.50	4340.500	-1.678	0.093
	At-risk	93.68	8618.50			
6	Non-at-risk	104.90	10909.50	4118.500	-2.579	0.010
	At-risk	91.27	8396.50			

Note. Only the tools showed significant differences were reported.
[a]$n = 104$. [b]$n = 92$.
*$p < .0083$.

Table 12: Results of the Mann Whitney U-Test to Compare the At-risk and Non-at-risk Groups' Daily Frequencies of Each Tool Use

Sequential Pattern Analyses

The previous study (Kang et al., 2017) performed sequential pattern mining using the cSPADE algorithm and identified students' learning behavior patterns of problem-solving and different behavior patterns of different performing groups. cSPADE mines constrained frequent sequences to discover sequences of actions among objects (e.g., students) in a given time period (Zaki, 2000, 2001). LSA is a statistical technique to examine whether a sequence of actions (i.e., an action follows another or the same action) in the overall behaviors of students achieves statistical significance (Bakeman & Gottman, 1997; Hou, 2015; Pohl et al., 2016). That is, LSA provides transitional probabilities indicating the likelihood that an action follows another or the same action. Then, a z-score of each transition is calculated to examine whether a transitional probability deviates significantly from its expected value (Bakeman & Gottman, 1997).

This process of analyses determines certain behavior transitions that occur significantly more often than others, which indicate significant behavior patterns.

The researcher extended the analysis of sequential behavior patterns using cSPADE with LSA, in order to propose an appropriate way of profiling students' behavior patterns using sequential gameplay data generated in the serious game. Table 13 shows each group's daily significant sequences with the adjusted residuals—calculated from LSA—and daily frequent sequential patterns—identified by cSPADE—by each group. A support value of cSPADE indicates that the percentage of students in a group used the sequence. A z score greater than 2.32 of LSA indicates the sequence in a group reaches a level of significance statistically ($p < .01$). Based on the significant sequences defined by LSA, transitional diagrams were created as shown in Figures 6-11. As described in the research contexts section, the game consists of ten in-game tools in a two-layer interface: the first layer including Alien Database, Probe Design Center, Mission Control Center, and Communication Center, and the second layer including Solar System Database, Missions Database, Concepts Database, Spectra, Periodic Table, and Notebook. Especially, the tools in the second layer can be overlaid anytime with any tool. In Figures 6-11, a square shape indicates a tool in the first layer, an octagon shape indicates a tool in the second layer, an arrow indicates the direction of a significant sequence, the number indicates z scores, and the line thickness indicates the level of significance.

Days	Non-at-risk Group		At-risk Group	
	LSA Results[a]	cSPADE Results[b]	LSA Results[a]	cSPADE Results[b]
Day1	periodic → spectra (z = 12.08) notebook → periodic (z = 7.22) missions → notebook (z = 5.97) spectra → concepts (z = 5.42) concepts → missions (z = 4.89) concepts → periodic (z = 4.85) concepts → notebook (z = 4.47) pdesign → mcontrol (z = 3.69) alien → alien (z = 3.45) communication → communication (z = 3.18) solar → communication (z = 2.67) solar → solar (z = 2.45)	solar (sup[c] = .96) alien (sup = .73) communication (sup = .60) alien → solar (sup = .52) communication → solar (sup = .42) mcontrol (sup = .38) solar → solar (sup = .30) solar → communication (sup = .30) alien → alien (sup = .28) communication → alien (sup = .25)	concepts → missions (z = 12.53) missions → notebook (z = 10.35) notebook → periodic (z = 9.59) periodic → spectra (z = 8.38) communication → communication (z = 5.74) alien → alien (z = 5.16) mcontrol → mcontrol (z = 4.76) pdesign → mcontrol (z = 4.30) notebook → notebook (z = 4.25) concepts → notebook (z = 4.06) pdesign → pdesign (z = 3.87) periodic → solar (z = 3.75)	solar (sup = .95) alien (sup = .66) communication (sup = .63) alien → solar (sup = .46) mcontrol (sup = .40) solar → solar (sup = .39) communication → solar (sup =.35) pdesign (sup = .30) missions (sup = .29) alien → communication (sup = .29) communication → alien (sup = .29) mcontrol → solar (sup = .27) alien → alien

Table 13 (continued)

			solar → spectra (z = 3.13) solar → solar (z = 3.01)	(sup = .26) periodic (sup = .25)
Day2	notebook → periodic (z = 6.75) concepts → missions (z = 6.54) communication → communication (z = 5.65) alien → alien (z = 4.98) missions → notebook (z = 4.81) periodic → spectra (z = 4.64) notebook → notebook (z = 3.83) pdesign → mcontrol (z = 3.28) spectra → spectra (z = 3.04) solar → solar (z = 2.64) concepts → periodic (z = 2.48)	solar (sup = .99) solar → solar (sup = .44) alien (sup = .35) communication (sup = .27)	concepts → missions (z = 11.64) notebook → periodic (z = 9.05) missions → notebook (z = 7.84) pdesign → pdesign (z = 5.84) periodic → spectra (z = 5.39) alien → alien (z = 5.32) communication → communication (z = 4.93) pdesign → mcontrol (z = 4.24) solar → solar (z = 3.40) solar → pdesign (z = 3.02) mcontrol → mcontrol (z = 2.51) spectra → mcontrol (z = 2.47)	solar (sup = .96) solar → solar (sup = .47) communication (sup = .37)

Table 13 (continued)

			concepts → notebook (z = 2.43)	
Day3	concepts → missions (z = 10.23) missions → notebook (z = 9.58) periodic → spectra (z = 6.72) pdesign → pdesign (z = 6.55) notebook → periodic (z = 5.97) pdesign → mcontrol (z = 5.86) communication → communication (z = 5.05) solar → solar (z = 3.03) solar → alien (z = 3.02) mcontrol → pdesign (z = 2.41)	solar (sup = .93) alien (sup = .43) solar → solar (sup = .36) pdesign (sup = .26)	notebook → periodic (z = 10.94) pdesign → mcontrol (z = 9.23) concepts → missions (z = 8.39) periodic → spectra (z = 8.04) missions → notebook (z = 6.92) solar → solar (z = 5.45) mcontrol → pdesign (z = 5.39) communication → communication (z = 5.19) spectra → spectra (z = 4.71) concepts → notebook (z = 4.4) mcontrol → mcontrol (z = 2.8) alien → alien (z = 2.58) concepts → periodic (z = 2.58)	solar (sup = .92) alien (sup = .43) solar → solar (sup = .40) communication (sup = .30) missions (sup = .26)

Table 13 (continued)

			notebook → concepts (z = 2.43)	
Day 4	pdesign → mcontrol (z = 10.52) missions → notebook (z = 9.65) periodic → spectra (z = 6.31) spectra → spectra (z = 5.64) communication → communication (z = 4.97) spectra → periodic (z = 4.91) pdesign → pdesign (z = 4.75) solar → alien (z = 4.61) concepts → missions (z = 4.29) mcontrol → mcontrol (z = 3.7) mcontrol → pdesign (z = 3.54) alien → spectra (z = 3.53) concepts → concepts (z = 3.39)	solar (sup = .83) alien (sup = .71) pdesign (sup = .57) communication (sup = .41) spectra (sup = .39) pdesign → pdesign (sup = .36) solar → alien (sup = .36) mcontrol (sup = .34) periodic (sup = .29) solar → solar (sup = .29)	missions → notebook (z = 9.9) pdesign → pdesign (z = 9.13) notebook → periodic (z = 8.78) communication → communication (z = 8.46) concepts → missions (z = 7) periodic → spectra (z = 5.92) mcontrol → mcontrol (z = 5.88) solar → solar (z = 5.57) concepts → notebook (z = 4.05) alien → alien (z = 3.75) pdesign → mcontrol (z = 3) spectra → spectra (z = 2.42)	solar (sup = .79) alien (sup = .54) pdesign (sup = .47) communication (sup = .31) solar → solar (sup = .28) missions (sup = .27) mcontrol (sup = .26) solar → alien (sup = .26)

Table 13 (continued)

	communication → alien (z = 3.24) mission → solar (z = 3.01) mcontrol → communication (z = 3) concepts → notebook (z = 2.92) solar → solar (z = 2.69)			
Day 5	concepts → missions (z = 8.17) pdesign → mcontrol (z = 7.96) periodic → spectra (z = 7.42) notebook → periodic (z = 7.36) missions → notebook (z = 7.19) concepts → notebook (z = 6.09) communication → communication (z = 5.63) mcontrol → mcontrol (z = 4.52) solar → alien (z = 3.69)	pdesign (sup = .65) alien (sup = .51) communication (sup = .48) solar (sup = .47) mcontrol (sup = .42) pdesign → mcontrol (sup = .35) pdesign → pdesign (sup = .33) mcontrol → pdesign (sup = .25)	notebook → periodic (z = 9.13) missions → notebook (z = 7.14) concepts → missions (z = 5.61) pdesign → mcontrol (z = 5.05) alien → alien (z = 4) solar → solar (z = 3.9) spectra → periodic (z = 3.86) communication → communication (z = 3.67) pdesign → pdesign (z = 3.65)	solar (sup = .62) pdesign (sup = .43) alien (sup = .37) communication (sup = .33)

Table 13 (continued)

	alien → spectra (z = 2.99) mcontrol → pdesign (z = 2.98) notebook → spectra (z = 2.93) periodic → solar (z = 2.91) mission → solar (z = 2.91) pdesign → pdesign (z = 2.88) spectra → notebook (z = 2.7) alien → concepts (z = 2.47)		periodic → spectra (z = 3.6) mcontrol → pdesign (z = 3.25)	
Day 6	missions → notebook (z = 13.19) concepts → missions (z = 11.59) periodic → spectra (z = 9.62) pdesign → mcontrol (z = 7.4) communication → communication (z = 7) pdesign → pdesign (z = 5.89)	pdesign (sup = .81) alien (sup = .58) communication (sup = .55) mcontrol (sup = .55) pdesign → pdesign (sup = .52) solar (sup = .45) pdesign → mcontrol	alien → alien (z = 11.72) concepts → missions (z = 11.39) periodic → spectra (z = 9.02) missions → notebook (z = 8.82) notebook → periodic (z = 8.41) pdesign → pdesign (z = 8.08) solar → solar	pdesign (sup = .58) solar (sup = .49) alien (sup = .44) communication (sup = .44) mcontrol (sup = .33) pdesign → pdesign (sup = .32)

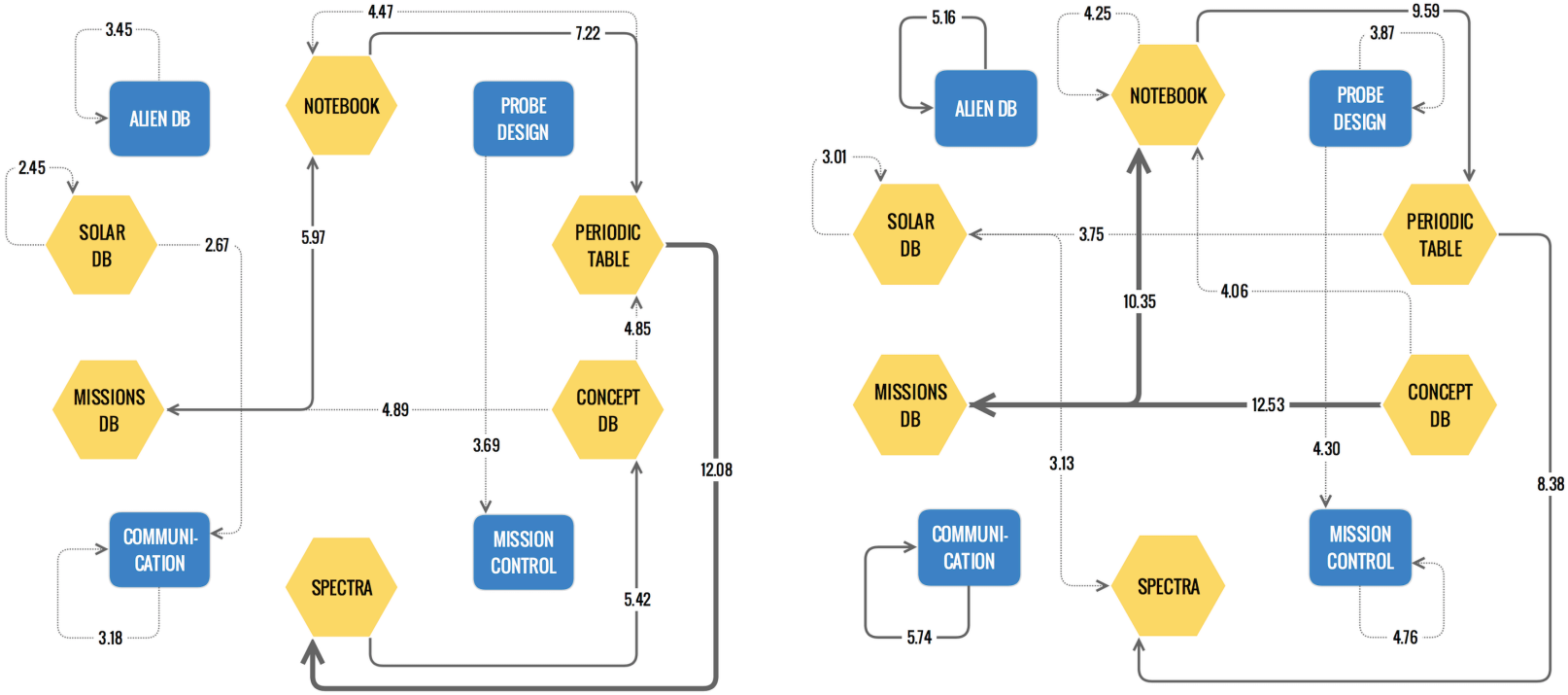
Table 13 (continued)

spectra → periodic	(sup = .45)	(z = 5.44)
(z = 5.32)	mcontrol → pdesign	mcontrol → mcontrol
alien → alien	(sup = .33)	(z = 4.62)
(z = 4.87)	pdesign → pdesign →	spectra → periodic
mission → solar	mcontrol	(z = 4.18)
(z = 4.4)	(sup = .32)	communication →
periodic → periodic	pdesign → communication	communication
(z = 3.86)	(sup = .32)	(z = 4.04)
concepts → notebook	communication → pdesign	concepts → notebook
(z = 3.42)	(sup = .31)	(z = 3.85)
mcontrol → communication	pdesign → pdesign →	concepts → periodic
(z = 3.21)	pdesign	(z = 3.85)
solar → alien	(sup = .31)	spectra → concepts
(z = 3.12)	mcontrol → communication	(z = 2.85)
concepts → periodic	(sup = .30)	
(z = 3.11)	mcontrol → mcontrol	
spectra → spectra	(sup = .29)	
(z = 3.04)	alien → pdesign	
communication → alien	(sup = .26)	
(z = 3)	mcontrol → pedisign →	
alien → spectra	mcontrol	
(z = 2.58)	(sup = .26)	
	solar → pdesign	
	(sup = .25)	
	pdesign → mcontrol →	
	communication	
	(sup = .25)	

Table 13 (continued)

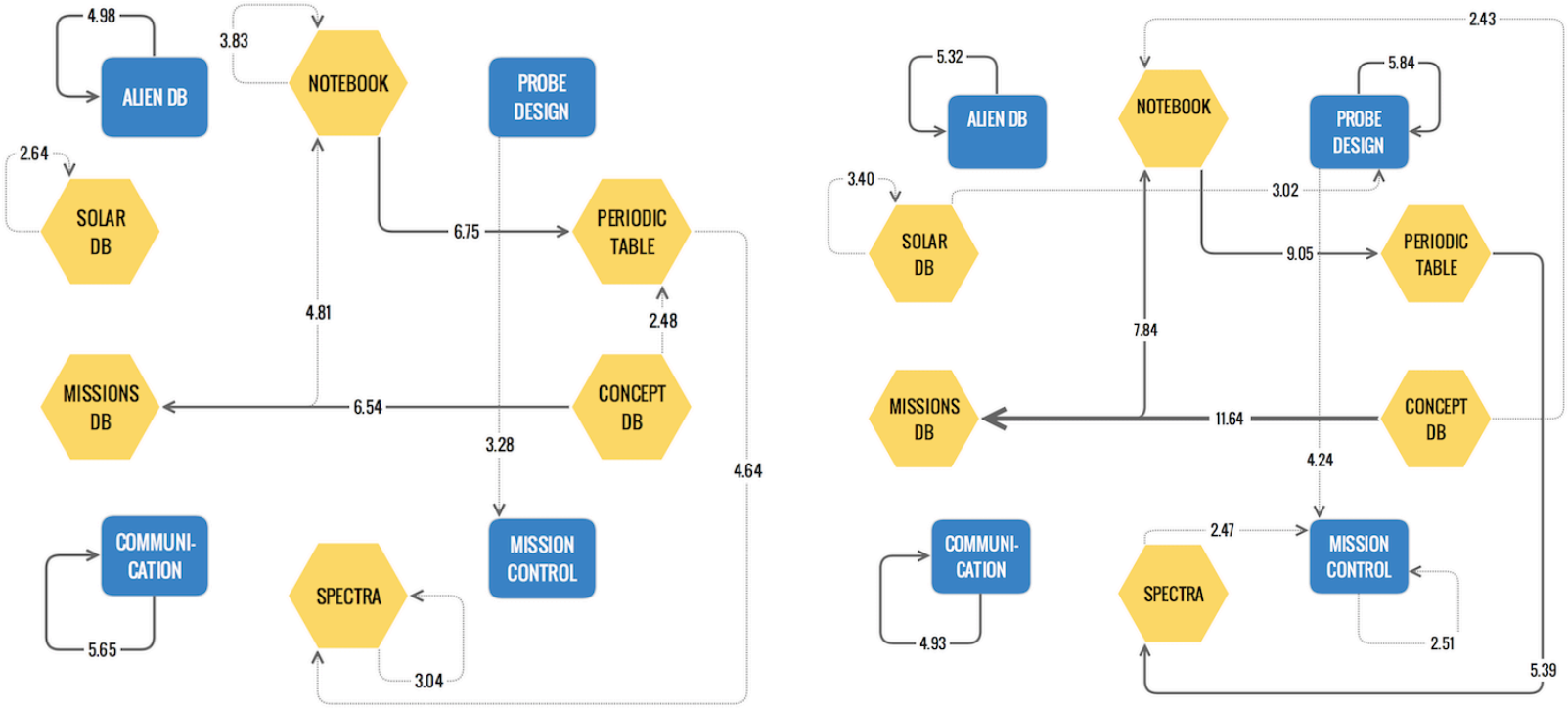
Note. [a]A Z-score identifying the probability that is higher than expected. Only the results showed above a *z*-score of 2.32 ($p < .01$; Bakeman & Gottman (1997)) were reported. [b]A set of frequent sequences mined ordered by its support value (maximum gap = 2, minimum support = .25). [c]A support value of a sequence of n-item (indicating a percentage of support value of students showed a sequence).

Table 13: Daily Frequent Patterns for Non-at-risk and At-risk Groups



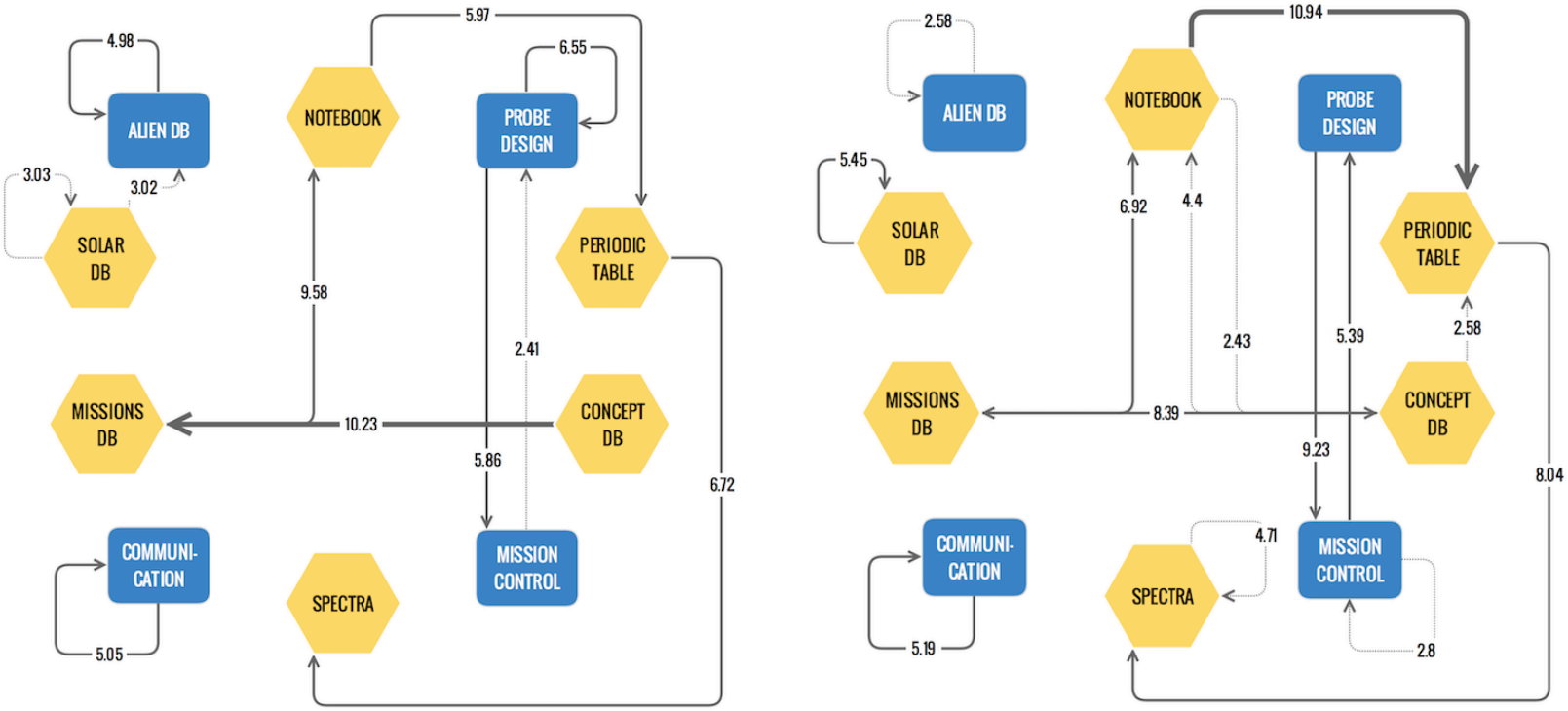
(a) Non-at-risk Students (n = 104) (b) At-risk Students (n = 92)

Figure 6: Navigational Transition Diagram of Day 1 ($p < .01$)



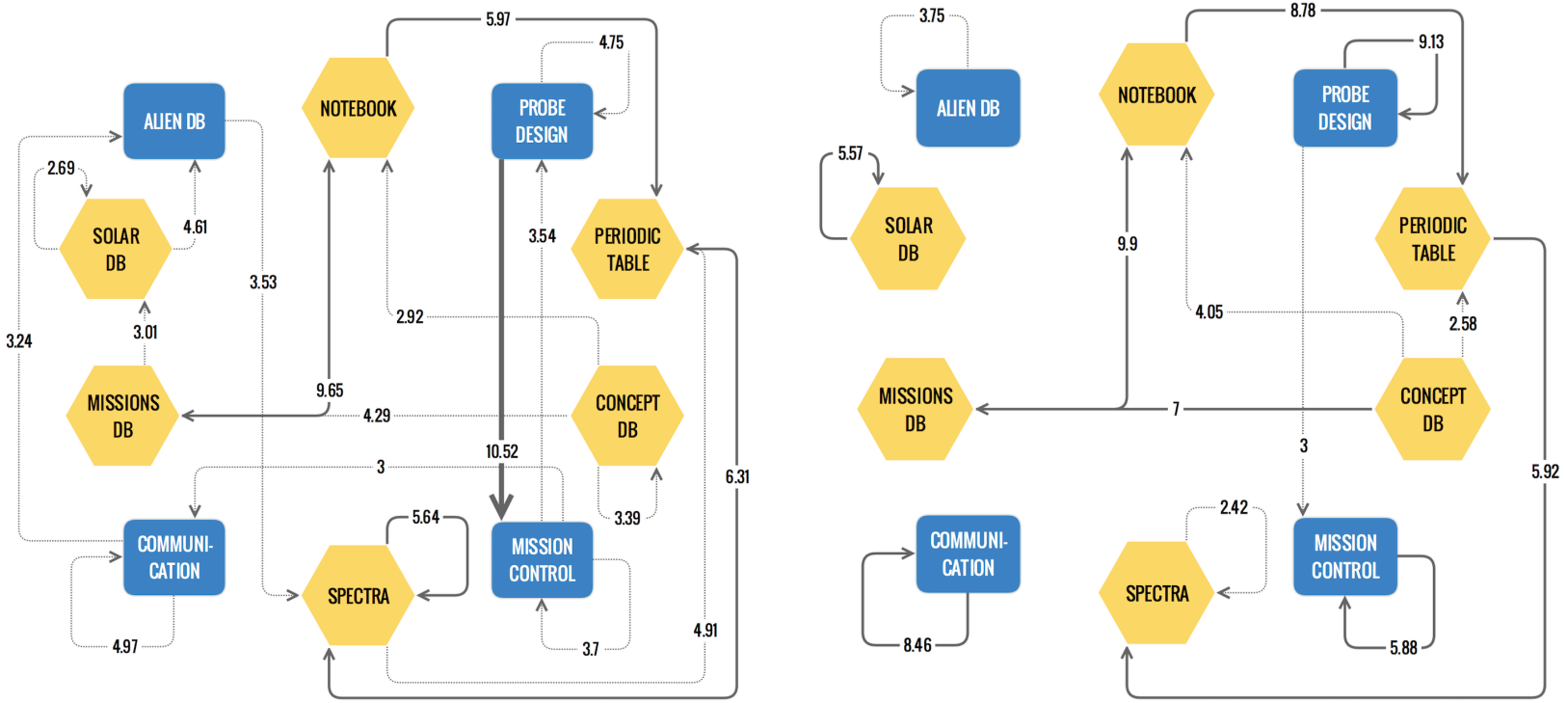
(a) Non-at-risk Students (n = 104) (b) At-risk Students (n = 92)

Figure 7: Navigational Transition Diagram of Day 2 ($p < .01$)



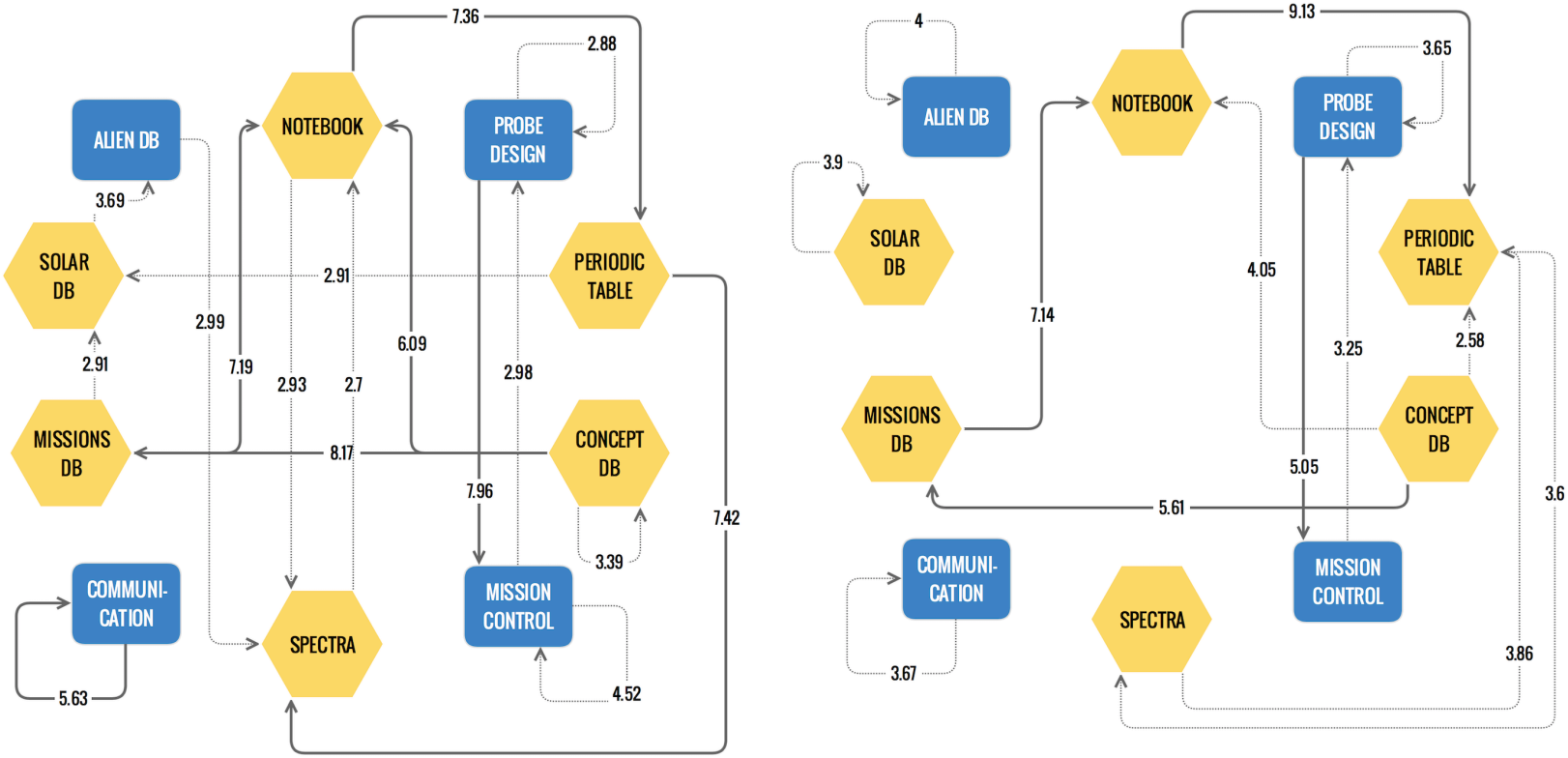
(a) Non-at-risk Students (n = 104) (b) At-risk Students (n = 92)

Figure 8: Navigational Transition Diagram of Day 3 ($p < .01$)



(a) Non-at-risk Students (n = 104) (b) At-risk Students (n = 92)

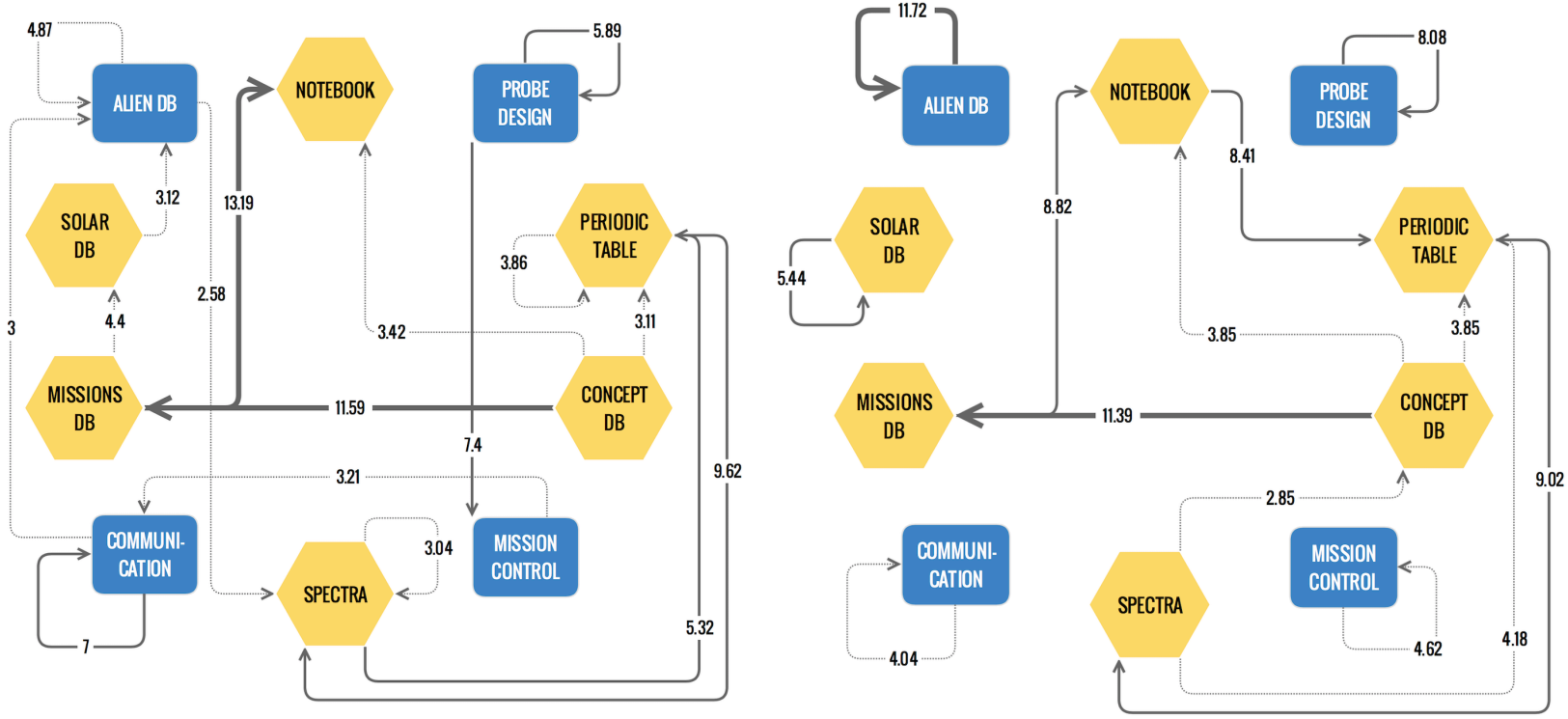
Figure 9: Navigational Transition Diagram of Day 4 ($p < .01$)



(a) Non-at-risk Students (n = 104)

(b) At-risk Students (n = 92)

Figure 10: Navigational Transition Diagram of Day 5 ($p < .01$)



(a) Non-at-risk Students (n = 104) (b) At-risk Students (n = 92)

Figure 11: Navigational Transition Diagram of Day 6 ($p < .01$)

On the first day, both non-at-risk and at-risk students tended to navigate the game environment by switching between different tools, which indicates these students attempted to understand how different tools can be used together. For example, the students frequently accessed Periodic Table followed by Spectra (periodic → spectra; $z_{\text{non-at-risk}} = 12.08$, $z_{\text{at-risk}} = 8.38$) or Notebook followed by Periodic Table (notebook → periodic; $z_{\text{non-at-risk}} = 7.22$, $z_{\text{at-risk}} = 9.59$). These sequences are not surprising as students would be expected to explore different tools and figure out closely related tools such as Periodic Table and Spectra, which would be necessarily used together to seek for a detailed spectrum of each chemical element during the problem-solving processes. On the other hand, the results of significant sequences of at-risk students showed that most of the significant sequences consist of a tool followed by the same tool (e.g., communication → communication). That is, the at-risk students had a tendency to use the same tools repeatedly. For example, the at-risk students showed more sequences consisting of the same tools (46.15%, 6 out of 14 significant sequences), compared to the non-at-risk students (25%, 3 out of 12 significant sequences). Furthermore, the sequential patterns in Figure 6 indicate that only the students in the non-at-risk group showed a significant bi-directional connection of 'Notebook after Periodic Table' and 'Periodic Table after Notebook,' while the at-risk group showed none bi-directional link. On the other hand, cSPADE found bi-directional transitions for both the non-at-risk (communication ↔ solar) and at-risk groups (communication ↔ alien). The results indicate both LSA and cSPADE found alien → alien, solar → communication, solar → solar as important sequences for the students in the non-at-risk group; however, the rest of significant sequences were not found in the cSPADE results. That is, although a sequence like periodic → spectra was not occurred very frequently in the non-at-risk group, the transitional probability of this sequence is higher than expected ($z = 12.08$).

On the second day, cSPADE found only a few frequent sequences for both groups including only one sequence of two items (solar → solar), which indicates almost the half of the students in each group (44% of non-at-risk students, 47% of at-risk students) used Solar System Database repeatedly. Similar to the LSA results of Day 1, both non-at-risk and at-risk students performed a certain degree of the learning behavior of switching different cognitive tools (i.e., periodic → spectra, notebook → periodic, missions → notebook, concepts → missions, pdesign → mcontrol). It is worth noting that the at-risk group was supplemented by additional sequences of pdesign → pdesign (z = 5.84), mcontrol → mcontrol (z = 2.51), and spectra → mcontrol (z = 2.47). In comparison, the non-at-risk group used Mission Control Center significantly only after Probe Design Center (z = 3.28). This suggests that the at-risk students were more prone to experiencing the tools which can be more appropriate to use during the later days of problem-solving processes. Since this is the second day, students are expected to access mainly Solar System Database and Alien Database to get an overview of the problem and gather information about planets and alien species (Kang et al., 2017; Liu et al., 2012).

On the third day, similar to Day 2, cSPADE discovered only a few frequent sequences for both non-at-risk and at-risk groups. The students in both groups still accessed Solar System Database the most frequently to find information about planets; however, a relatively small number of students, compared to Day 2, showed a frequent sequence of solar → solar (36% of non-at-risk students, 40% of at-risk students). Interestingly, the non-at-risk students showed an additional significant sequence of solar → alien (z = 3.02), indicating these students began matching planets with alien needs to find suitable places for each alien, while the at-risk students did not show any transition between these two tools: Solar System Database and Alien Database, but showed only the sequences consisting of the same tools (i.e., solar → solar, alien → alien). Another

noticeable finding is that both groups had the unidirectional sequential pattern of pdesign → mcontrol for the first two days, and then used the tools in the reverse way (i.e., mcontrol → pdesign) on the third day (see Figure 8). This interaction suggests that the students first explored these tools at the early days, and then tended to gather additional information about planets or moons in our solar system that they needed to confirm a suitable home for the aliens. For example, they designed and launched probes in Probe Design Center and then opened Mission Control Center to view the gathered results from the probes they launched. The students either received the information they needed to confirm the potential homes for the aliens, or received error messages they needed to interpret why their probes failed. After this exploring phase, the students learned how to design appropriate probes, and then they revisited Probe Design Center to remedy errors from previous probe and gather new information about the planets that was not gathered with previous probes (Liu et al., 2009).

From Day 4, the results of both cSPADE and LSA (see Figure 9) show the non-at-risk students had more sequential links between different tools, while the at-risk students had the most probable cases of those where one tool follows the same tool. This indicates the non-at-risk students became more actively involved in the process of problem-solving, while the at-risk students had a tendency to do the same things repeatedly. For example, approximately 40% of the non-at-risk students continuously accessed Solar System Database and Alien Database in order to gather the necessary information, which can be used to confirm good choices of planets or eliminate bad planet choices. The sequence of alien → spectra ($z = 3.53$) indicates that the non-at-risk students figured out Alien Database included incomplete information of the required chemical elements for some aliens (e.g., providing only a picture of spectrum of the element without the name of the element), which can be found in Spectra. These students also had the bi-directional

sequential pattern of periodic ↔ spectra, which can be shown as they compared the chemical elements' information of planet and moons with the aliens' need. Previously, the Mann Whitney U-test revealed the significant differences of the frequencies of four in-game tools between two groups: Alien Database, Probe Design Center, Solar System Database, and Spectra. Especially, the mean frequency differences of Solar System Database and Spectra between two groups achieved the level of significance only on Day 4. Together with the significant sequences discovered from cSPADE and LSA, these results suggest the most critical tools to make a progress of problem-solving before students generate a hypothesis of testing a suitable home for the aliens: Alien Database, Solar System Database, Spectra, Periodic Table, Probe Design Center, and Mission Control Center. Additionally, the LSA results showed the interesting sequences related to Communication Center, which allows students to select a suitable planet for each alien species, write a rationale for their choice of alien habitat, and then submit the recommendation. The transition from Mission Control Center to Communication Center (i.e., mcontrol → communication, $z = 3$) indicates the non-at-risk students confirmed that a planet met all of alien's needs and then accessed Communication Center to write a recommendation. Classroom observation revealed that the teachers randomly assigned one alien species to each student. Once the student successfully sent the alien to the possible homes, the teacher assigned another alien for the student to get additional points. Therefore, the sequence, communication → alien ($z = 3.24$), suggests the non-at-risk students began to research about new alien after submitting the final solution of the previous alien.

The cSPADE results of Day 5 show the at-risk students had none frequent sequences of two items, while the non-at-risk students had more active transitions between Probe Design Center and Mission Control Center (i.e., pdesign → mcontrol,

pdesign → pdesign, mcontrol → pdesign). The most interesting finding of Day 5 is the non-at-risk students' solar → solar sequence did not reach a level of significance. This finding suggests most of these students acquired enough information from Solar System Database and therefore did not necessarily revisit the tool. A sequence of periodic → solar (z = 2.91) is also the unique transition existing only in the non-at-risk students' significant sequences. Although the at-risk students showed the same sequence on Day 1, it was not considered as a meaningful behavior since students were expected to explore different tools without any strategies on the first day. In addition, similar to Day 4, the at-risk group did not show any significant transition between Alien Database and any other tools, but only the transition from Alien Database to Alien Database (alien → alien, z = 4.00), indicating these students did not make any further progress of matching aliens' needs and other information provided in the rest of tools such as Solar System Database.

Interestingly, cSPADE yielded the highest number of frequent sequences including the sequences of the three items for the non-at-risk group on the last day (*minsup* > .25). This indicates, compared with the previous days, more students in this group accessed the tools in similar sequences more frequently during this day. Since the fifth day, the non-at-risk students continuously accessed Probe Design Center and Mission Control Center to integrate all gathered information to confirm their recommendations, and then visited Communication Center to submit their final solutions. In contrast, the at-risk students (approximately 30%) had only one frequent sequence of two items (pdesign → pdesign). Unlike the previous days, the non-at-risk students had the unidirectional sequence, pdesign → mcontrol (z = 7.4) on the last day. This finding suggests, through trial-and-error during the previous days, these students learned how to remedy the errors from their probes, and therefore, they were less likely to return to Probe Design Center to send multiple probes. The sequence, mcontrol → communication

(z = 3.21), indicates these students accessed Communication Center after they gathered more information about the planets from Mission Control Center. Furthermore, the non-at-risk students had the significant sequences of communication → alien and mcontrol → communication on Day 4 and Day 6, and periodic → solar on Day 5. These patterns suggest that, as discussed on Day 4, the non-at-risk students most likely gathered additional information about the first assigned alien species during the process of writing a recommendation, or submitted at least one recommendation of a good home for the first alien and gather information about new species (i.e., communication → alien). On Day 5, the non-at-risk students were still in the process of gathering the new information across different databases such as Solar System Database and Periodic Table (i.e., periodic → solar), which is expected since they needed additional information to eliminate choices of worlds for the second alien or determine second possible world for the first alien. Then, finally they attempted to submit a solution(s) (e.g., communication → alien, mcontrol → communication) on the last day.

Summary of Analyses on Identifying Navigation Behaviors

The purpose of the first research question was to examine what differences exist between at-risk and non-at-risk students' navigation behavior patterns. First, the researcher investigated the differences in the posttest performance after adjusting for the pretest scores between at-risk and non-at-risk groups using One-Way ANCOVA. The results showed the non-at-risk group had significantly higher improvement on science knowledge than the at-risk group. However, the effect size of at-risk factor is small (partial eta squared (η_p^2) = .081) by the rule of thumb (Cohen, 2013; MRC Cognition and Brain Sciences Unit, 2009).

Considering the previous result, the researcher further examined what navigation behavior patterns emerged for students in the at-risk and non-at-risk groups during their problem-solving process. The researcher first adapted a Mann Whitney U-test to examine whether the daily frequencies of each in-game tool use differed between at-risk and non-at-risk students. The results showed that there were significant differences between the rank averages of the groups' frequencies of tool use during the later days (i.e., Days 4-6); that is, the students in the non-at-risk group tended to access overall in-game tools more often than those in the at-risk group. In addition to the U-test, the researcher integrated LSA and cSPADE to identify the two groups' navigation behavior patterns in this game context. The results from both methods revealed that problem-solving strategies were differently used between the non-at-risk and at-risk students within this environment through the six days of their gameplay period. During the first day, both groups showed a similar tendency to explore most of the in-game tools. On the second day, the at-risk group was prone to use more tools repeatedly (i.e., pdesign → pdesign, mcontol → mcontrol), while the other group showed the transitional behavior between the tools (i.e., pdesign → mcontol). The significant patterns discovered over the later days suggest that the at-risk group continued to revisit the same tools, which means these students seemed not to understand how different tools could be used together. On the other hand, the non-at-risk group showed transitions between different tools, which indicates that they became more strategic in their problem-solving processes. Together with the U-test results, the sequential analyses helped to explore students' various navigation behavior patterns as well as discover the different problem-solving processes between the non-at-risk and at-risk students.

EFFECT ON SCIENCE KNOWLEDGE

This study is further interested in understanding students' behaviors in Probe Design Center that supports students' scientific inquiry process by generating and refining hypotheses and designing their probes. Five game metrics were proposed to understand students' scientific inquiry behaviors in the game. Two research questions were asked to examine the effect of students' inquiry behaviors on science knowledge and to identify students' inquiry behavior patterns:

3) What is the relationship between students' inquiry behaviors in Probe Design Center and their learning performance?
4) What scientific inquiry behavior patterns emerge as students engage with Probe Design Center?

Research Question 3: What is the relationship between students' inquiry behaviors in Probe Design Center and their learning performance?

Two hierarchical regression analyses were carried out to investigate how much extra variation in students' learning performance (i.e., average solution scores for the first analysis and SSKT posttest scores for the second analysis) can be explained by the addition of scientific inquiry behavior variables generated from Probe Design Center. The predictors are the four behavior variables (i.e., number of repeated trials, amount of new information, amount of redundant information, and number of errors), and the covariates are the number of launched probes variable and the school variable (School A, School B). As mentioned in the participants section, the classroom observations in School A revealed their limited usage of Probe Design Center; therefore, the researcher included additional data from School B (see Table 2) to build a model controlling for a school.

Prior to performing the regression analyses, the data were evaluated for violation of assumptions. First, the linear relationships between the dependent variable and each of the predictors were examined by plotting residuals against each predictor and the residuals against the predicted values. The plots did not show any systematic pattern or clustering of the residuals, indicating no violations in linearity (Stevens, 2009). Homoscedasticity was checked by visual examination of the plots of the standardized residuals by each predictor and the regression standardized predicted value. The residuals were randomly scattered around zero indicating even distribution (Osborne & Waters, 2002). The multicollinearity was examined by VIF (i.e., VIFs < 10), an index of the amount that the variance of each regression coefficient is increased over that with uncorrelated independent variables (Keith, 2006). The boxplot of the residuals and Durbin-Watson statistic (i.e., 1.765 for the first analysis and 2.318 for the second analysis) confirmed the independence of the residuals, indicating that errors associated with one observation were not correlated with the errors of any other observation. The ordered leverage values indicated the two cases that had the values above .2. Therefore, the two cases were removed to run a regression analysis. Also, there was no Cook's Distance values above 1, which indicates there is no cases that are influential. (Cook & Weisberg, 1982). Lastly, the normal P-P plots of regression standardized residuals indicated that the assumption of normality was not violated. Table 14 shows the descriptive statistics of each variable.

Variable	*M*	*SD*
Number of errors	1.92	2.417
Number of repeated trials	.48	.734
Amount of new information	4.76	4.404
Amount of redundant information	.509	.611
Number of launched probes	2.03	2.549
Average solution score	1.96	1.83
SSKT posttest score	65.61	17.04

Table 14: Basic Descriptive Statistics of Variables ($N = 133$)

Specifically, the first hierarchical multiple regression was conducted to determine if the addition of four scientific inquiry behavior variables generated from Probe Design Center improved the prediction of in-game learning performance (i.e., average solution scores) over and above a school and number of launched probes alone. The covariates were entered first in the regression equation, and then the four independent variables of interest were entered into the equation. Table 15 shows the details on the first regression models. The full model of five game metrics and school to predict an average solution score was statistically significant, $R^2 = .188$, $F(4, 126) = 4.859$, $p < .01$; adjusted $R^2 = .149$. The addition of weight to the prediction of average solution score (Model 2) led to a statistically significant increase in R^2 of .081.

Together, the predictors explained about 14.9% of the variability in the average solution scores. A student's predicted posttest score is equal to $0.771 + 0.377\times$

$(School) + 0.023\times(Probes) - 0.010\times(Errors) + 0.135\times(Repeated\ Trial) + 0.151\times(New\ Information) + 0.046\times(Redundant\ Information)$, in which the school variable is 1 (School A) or 0 (School B) and the last five variables are the game metrics. The standardized beta coefficient indicates the new information ($\beta = .365$, $p < .01$) was significantly and positively related to the average solution scores. Although the other three game metrics did not contribute to the full regression model, all predictors were included to see if this prediction is related to the criterion after controlling for all the other predictors in the model. In general, the more a student receives new information, the higher solution scores they will get from the solution submission on average, after controlling for the other variables in the model. In other words, for every additional amount of new information, a student would be predicted to have approximately 0.365 higher points on the average solution score.

	In-game Performance (Average Solution Score)			
	Model 1		Model 2	
Variable	*B*	*β*	*B*	*β*
Constant	.781**		.771	
School	.422	.113	.377	.101
Number of launched probes	.222**	.350	.023	.037
Number of errors			-.010	-.013
Number of repeated trials			.135	.054
Amount of new information			.151**	.365
Amount of redundant information			.046	.064
R^2	0.107		0.188	
F	7.815**		4.859**	
ΔR^2	0.107		0.081	
ΔF	7.815**		3.126**	

Note. *$p < .05$; **$p < .01$; B: unstandardized regression coefficient; *β*: standardized coefficient

Table 15: Hierarchical Multiple Regression Analysis for Variables Predicting Average Solution Score ($N = 133$)

Table 16 shows the details on the second set of regression models. The full model of five game metrics and school to predict a SSKT posttest score was statistically significant, $R^2 = .299$, $F(4, 126) = 8.941$, $p < .01$; adjusted $R^2 = .265$. Specifically, the addition of weight to the prediction of SSKT posttest score (Model 2) led to a statistically

significant increase in R^2 of .058. Together, the predictors explained about 26.5% of the variability in the SSKT posttest scores. A student's predicted posttest score is equal to $64.625 - 9.812\times(School) + 2.53\times(Probes) - 0.727\times(Errors) - 2.57\times(Repeated\ Trial) + 0.423\times(New\ Information) - 1.392\times(Redundant\ Information)$, in which the school variable is 1 (School A) or 0 (School B) and the last five variables are the game metrics. The standardized beta coefficients indicate the school classification (β = -.281, p < .01) and the amount of redundant information (β = -.208, p < .01) were significantly and negatively related to the SSKT posttest scores. The number of launched probes (β = .428, p < .01) was positively related to the posttest scores. All predictors were included to see if each predictor is related to the criterion after controlling for all the other predictors in the model. In specific, for every one difference on the number of launched probes, a student would be predicted to have approximately 2.53 higher points on the posttest score. An increase in the redundant information of 1 is associated with a decrease in the posttest score of 1.392 points. The school variable was coded as: 1 = School A and 0 = School B. Therefore, all other things being equal, students in School B have the posttest scores that are 9.812 points greater than students in School A.

	Post-game Performance (SSKT Posttest Score)			
	Model 1		Model 2	
Variable	*B*	*β*	*B*	*β*
Constant	67.382**		64.625**	
School	-12.309**	-.353	-9.812**	-.281
Number of launched probes	1.406**	.238	2.53**	.428
Number of errors			-.727	-.104
Number of repeated trials			-2.573	-.111
Amount of new information			.423	.109
Amount of redundant information			-1.392*	-.208
R^2	0.241		0.299	
F	20.587**		8.941**	
ΔR^2	0.241		0.058	
ΔF	20.587**		2.610**	

Note. *$p < .05$; **$p < .01$; B: unstandardized regression coefficient; β: standardized coefficient

Table 16: Hierarchical Multiple Regression Analysis for Variables Predicting SSKT Posttest Score ($N = 133$)

Research Question 4: What scientific inquiry behavior patterns emerge as students engage with Probe Design Center in the serious game Alien Rescue?

In the previous analysis, two hierarchical regression analyses were performed to investigate how much extra variation in students' learning performance can be explained

by the addition of scientific inquiry behavior variables generated from Probe Design Center. The analyses confirmed the addition of scientific inquiry behavior variables to the prediction of both in-game and after-game learning performance (i.e., average solution scores and SSKT posttest scores) led to statistically significant increases in the scores. For this research question, the researcher further conducted cluster analyses with five game metrics to discover any distinctive inquiry behavior groups in each school and examined the characteristics of each group. Table 17 shows the basic descriptive statistics of students' game metrics and learning performance in each school.

Variable	School A (n = 82)		School B (n = 51)	
	M	*SD*	*M*	*SD*
Number of errors	1.74	2.38	2.22	2.49
Number of repeated trials	0.43	0.67	0.57	0.83
Amount of new information	3.87	3.85	6.2	4.88
Amount of redundant Information	1.99	2.58	2.1	2.53
Number of launched probes	3.33	2.60	5.43	2.85
Number of saved aliens[a]	.99	.98	2.84	2.60
Average solution score	1.94	1.91	1.99	1.69
Total solution score[b]	3.06	3.63	9.20	9.74
SSKT Posttest score	59.76	17.53	75.02	11.03

Table 17 (continued)

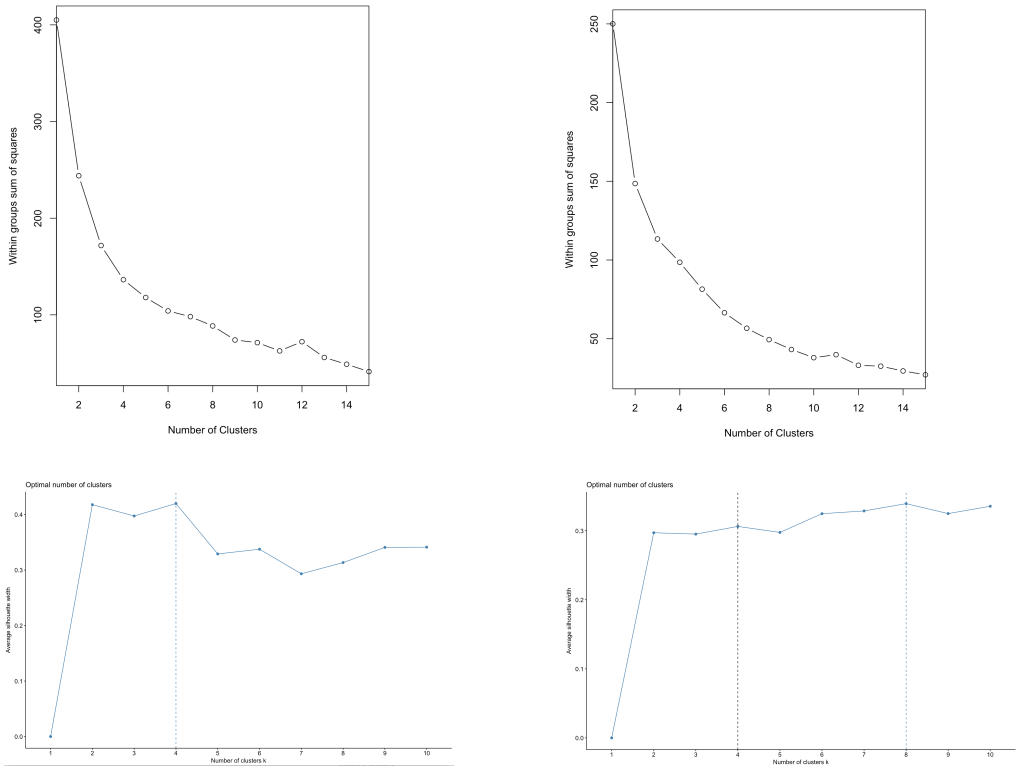
Note. [a]The number of aliens, for which students submitted at least one solution; [b]Sum of all submitted solution scores

Table 17: Basic Descriptive Statistics of Game Metrics and Learning Performance for Each School

Each variable was standardized before the cluster analyses to deal with a different scale of each variable. A partitioning method such as k-means requires a researcher to specify the numbers of clusters to be generated. The numbers of clusters can be subjective depending on a method used for measuring similarities or parameters used for partitioning (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). Therefore, to inspect the optimal number of clusters for each school, this study used two different methods: Ward method and the average silhouette. The Ward method minimizes the total within-cluster variance. This method finds the pair of clusters that leads to minimum increase in total within-cluster variance after each step of merging (Charrad et al., 2014). The average silhouette measures the quality of a clustering, which determines how well each object falls in its cluster by computing the average silhouette width of observations for different numbers of clusters. A high average silhouette width indicates a good clustering (Kaufman & Rousseeuw, 1990).

As shown in Figure 12, an elbow in each graph (Stevens, 2009) indicates that the number of clusters that the elbow indicates can minimize the distance between cases within each cluster and maximize the distance between clusters, as well. Although the average silhouette graphs recommended 6 and 8 clusters for School A and B, the average silhouette widths were not much different from the ones near the recommended number of clusters. In addition, the researcher conducted several cluster analyses by increasing

the number of clusters from two to eight, as suggested in the within groups sum of squares results. The results indicated a bigger number of groups yielded a group with only a few students and additional clusters merely split the objects into the predefined number of cluster without additional interpretive value. Therefore, four clusters were determined in consideration of different perspectives such as the within groups sum of squares and average silhouette graphs for each school.



(a) School A (b) School B

Figure 12: Within groups sum of squares (above) and Average Silhouette by number of clusters (below)

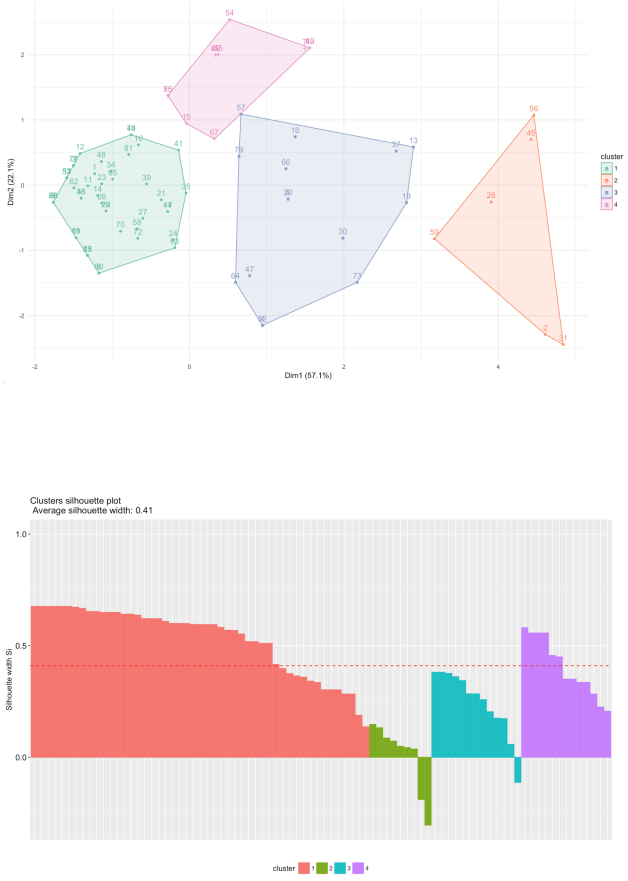
Note. Black dotted line is the determined number of clusters; Blue dotted line is the suggested number of clusters.

This study conducted a *k*-medoids clustering using *pam* algorithm on the derived cluster number for each school. Then, the clustering results were examined if there were differences in the scientific inquiry behaviors (i.e., five game metrics) between four clusters of students in each school. The researcher conducted a non-parametric test, a Kruskal-Wallis H test, since the data failed the major assumption of the one-way ANOVA, the non-normally distribution assumption.

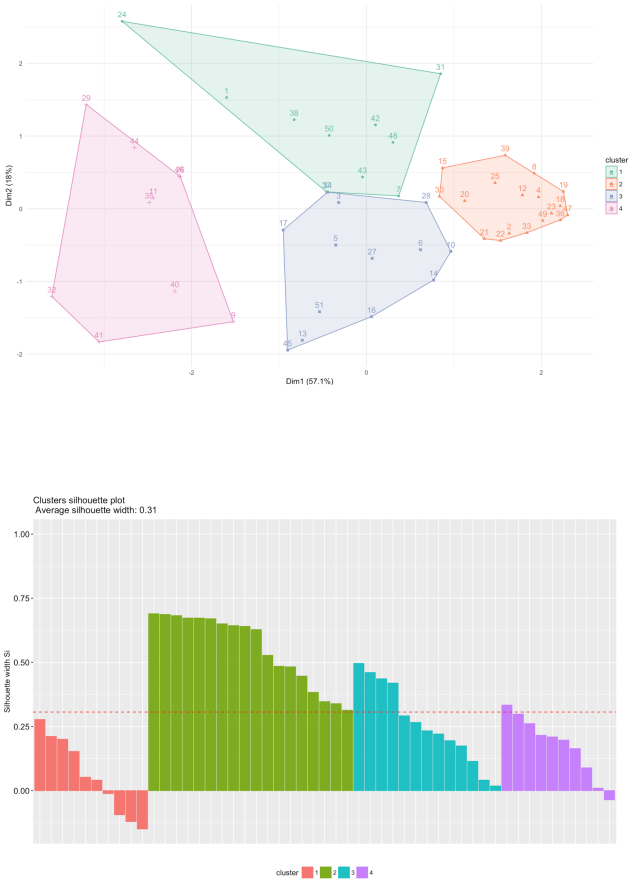
Figure 13 showed the results of cluster analyses including the cluster plots and average silhouette plots for each school. Since k-medoids applies a dimension reduction algorithm to partitioning data to the given number of clusters, it produces a two-dimensional plot for indicating the partitioning results (see Figure 13). Overall, the clusters do not show any significant ambiguity between different clusters; that is, the cluster plots show the most of members belong to a certain cluster. The silhouette plots provide more detail information of the partitioning results by calculating a silhouette value of each member in each cluster and the average silhouette value of each cluster. The silhouette value ranges from -1 to +1, in which a high silhouette value indicates that a member is well assigned to its own cluster, and a low value indicating poorly assigned to neighboring clusters. Figure 13 shows a few negative silhouette values. For example, two negative values of Cluster 2 in School A and four negative values of Cluster 1 in School B. However, the researcher concluded these clustering solutions are appropriate as most members show high silhouette values (Kaufman & Rousseeuw, 1990).

Table 18 shows that the average values and mean ranks of the scientific inquiry behaviors (i.e., five game metrics) as exhibited by the four clusters of students in each

school achieved the level of significance (χ^2), which indicates the students were well-partitioned into each group. The four clusters of School A comprise 49, 9, 13, and 13 students, respectively, accounting for 58.33%, 10.71%, 15.48%, and 15.48% of the total students. The four clusters of School B comprise 10 (19.61%), 18 (35.29%), 13 (25.49%), and 10 (19.61%) students. This study further examined the background of the four clusters of students in each school by indicating the average posttest SSKT scores, average solution scores, and the percentage of students who submitted at least one solution.



(a) School A



(b) School B

Figure 13: Cluster Analyses Results: Cluster plots (up) and Silhouette plots (down).

Schools	Clusters	Number of students with solution(s) submitted[a]	Average posttest score	Metrics (Mean rank)				
				Repeated Trials	Errors	New Information	Redundant Information	Launched Probes
School A	Cluster 1 (n = 49, 59.76%)	26 (53.06%)	56.63	0.10 (31.38)	1.04 (35.39)	1.65 (28.27)	0.88 (30.89)	1.61 (25.59)
	Cluster 2 (n = 7, 8.54%)	6 (85.71%)	61.90	1.57 (71.71)	5.86 (71.93)	7.00 (63.50)	9.00 (79.00)	8.29 (75.79)
	Cluster 3 (n = 13, 15.85%)	12 (92.31%)	62.50	1.15 (65.59)	3.69 (63.08)	4.00 (46.65)	1.92 (47.50)	6.23 (68.00)
	Cluster 4 (n = 13, 15.85%)	13 (100.00%)	67.63	0.31 (39.19)	0.23 (24.54)	10.38 (74.38)	2.46 (55.31)	4.23 (56.50)
	χ^2			48.43***	35.35***	47.56***	34.67***	60.58***
School B	Cluster 1 (n = 10, 19.61%)	6 (60.00%)	76.52	1.00 (35.4)	4.60 (38.45)	4.20 (20.85)	1.60 (26.6)	6.50 (31.6)
	Cluster 2 (n = 18, 35.29%)	10 (55.56%)	69.08	0.17 (19.67)	0.67 (15.92)	1.67 (11.83)	0.44 (15.42)	2.39 (10.14)
	Cluster 3 (n = 13, 25.49%)	10 (76.92%)	81.94	0.08 (17.69)	1.15 (20.62)	9.08 (35.65)	1.85 (25.81)	6.62 (32.35)

Table 18 (continued)

Cluster 4	8 (80.00%)	75.22	1.50	4.00	12.60	5.90	8.30
(n = 10, 19.61%)			(38.8)	(38.7)	(44.1)	(44.7)	(40.7)
χ^2			24.72***	25.58***	38.12***	26.77***	34.52***

Note. [a]The number of students who submitted at least one solution (Percentage of the students in each group); ***p<0.001

Table 18: Cluster analysis results of the students' scientific inquiry behavior for each school.

To further interrogate potential patterns between the metrics, radar plots—as the representation of behavior patterns between different groups—were derived as shown in Figures 14 and 15. Since each metric has a different scale, this study used the mean rank of each metric to visualize each group's behavior patterns. Along with the analysis results (see Table 18), the radar plots (see Figures 14-15) indicated that the students' inquiry behavior patterns in this game were distinctively different.

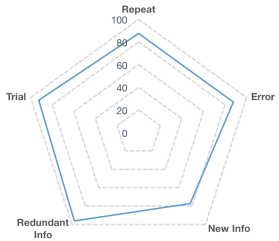
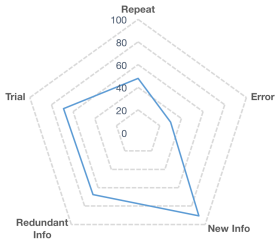
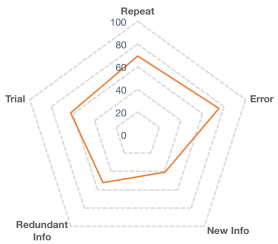
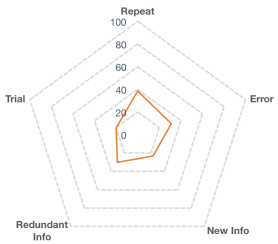
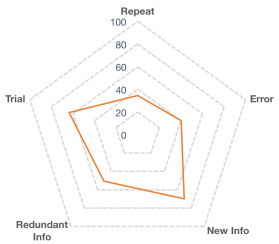
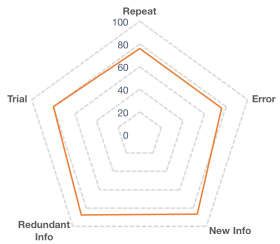
				
Saved aliens[a]	0.61	1.71	1.31	1.69
Total solution[b]	1.76	6.43	3.46	5.77
Avg. solution[c]	1.31	3.29	2.15	3.38
Posttest[d]	56.63	61.90	62.50	67.63
	(a) Cluster 1 (n = 49)	(b) Cluster 2 (n = 7)	(c) Cluster 3 (n = 13)	(d) Cluster 4 (n = 13)

Figure 14: Radar Plots of Cluster Analysis Results and Students' Learning Performance of School A

Note. All mean ranks were converted into a percent for the radar plots. [a]Average of the numbers of aliens, for which students in each group submitted at least one solution; [b]Average of the total solution scores of students in each group; [c]Average of the average solution scores of students in each group; [d]Average of the SSKT posttest scores of students in each group.

				
Saved aliens	2.00	2.50	3.46	3.50
Total solution	5.30	7.61	11.38	13.10
Avg. solution	1.35	1.50	2.40	2.96
Posttest	76.52	69.08	81.94	75.22

(a) Cluster 1 (n = 10) (b) Cluster 2 (n = 18) (c) Cluster 3 (n = 13) (d) Cluster 4 (n = 10)

Figure 15: Radar Plots of Cluster Analysis Results and Students' Learning Performance of School B

As for School A, approximately 60% of the students are centered in Cluster 1, and the mean ranks of five inquiry behaviors were lower than those of the other three clusters. That is, more than half of the students in School A did access Probe Design Center infrequently compared with the rest of clusters. This group showed the lowest solution submission rate; that is, 46.4% of the students in this group did not submit any solution, and this group saved less than one alien (i.e., 0.61) on average during the entire gameplay. In addition, these students achieved the lowest solution scores and SSKT posttest scores than the other three clusters.

Only seven students (9%) are centered in Cluster 2; their mean ranks of inquiry behaviors were higher than the other three clusters. In terms of the solution submission rate, approximately 86% of the students in this group submitted at least one solution, and the students saved one or two aliens (i.e., 1.71) on average during the entire gameplay. Therefore, approximately 10% of the students playing this game were prone to experiencing Probe Design Center (i.e., the highest number of launched probes). In specific, this group received relatively lower amount of new information compared with other metrics—the highest ranks of the amount of redundant information, the number of errors, the number of repeated trials, and the number of launched probes. These findings suggest this group tend to adopt a trial-and-error approach by launching many probes. The students in this group achieved relatively higher in-game performance ($Solution_{Average}$ = 3.29; $Solution_{Total}$ = 6.43); however, lower after-game performance (Posttest = 61.90), compared with other three groups. Together, these findings suggest that the trial-and-error approach may positively affect their learning performance, but it

may require more time. That is, the students obtained additional information that they needed to recommend possible homes for aliens; however, they did not achieve enough knowledge required for successful after-game performance.

About 16% of the students in School A (n = 13, 15.85%) are in Cluster 3. The students exhibited the average inquiry behavior regarding all five metrics; that is, their overall mean rank of each metric lies between Cluster 1 and Cluster 2. This group showed relatively lower ranks of the amount of new and redundant information, compared with the other metrics. In terms of the solution submission rate, approximately 92% of the students in this group submitted at least one solution, and the students saved one or two aliens (i.e., 1.31) on average during the entire gameplay. The students in Cluster 3 performed lower ($Solution_{Average}$ = 2.15; $Solution_{Total}$ = 3.46) than the students in Cluster 2 ($Solution_{Average}$ = 3.29; $Solution_{Total}$ = 6.43) at their in-game performance. However, these students achieved their SSKT posttest scores slightly above the mean score of School A (Posttest = 62.50), which is similar to Cluster 2. The results of the regression analyses showed the amount of new information as a significant variable in prediction of students' in-game performance. In support of the regression analyses, these findings showed the amount of new information is positively related to students' in-game performance.

Lastly, the rest of the students (n = 13, 15.85%) are centered in Cluster 4. Every student in Cluster 4 was able to submit at least one solution during the entire gameplay period. Overall, these students showed the highest learning performance from both in-game and after-game performance scores ($Solution_{Average}$ = 3.38; Posttest = 67.63),

among all groups. As shown in Figure 15, the radar plot of this group showed a spike on the new information. It is worth noting that these students launched approximately four probes—slightly above the average of School A (i.e., 3.33)—but they received the highest amount of new information compared with the other groups. In addition, this group got the second highest mean rank of the amount of redundant information (mean rank = 55.31), which was between Cluster 2 (mean rank = 79.00) and Cluster 3 (mean rank = 47.50). Furthermore, their mean ranks of repeated trials and errors are relatively lower than the other groups. That is, their repeated trials and errors ranked the third with 39.19 and 24.54, followed by Cluster 1 with 31.38 and 35.39.

The cluster results of School B also showed different inquiry behavior patterns among four clusters. First, approximately 20% of the students are centered in Cluster 1. These students exhibited the mean ranks of all five metrics lay between other clusters, which is similar with Cluster 3 of School A. Similar to Cluster 3 of School A, these students performed relatively lower than other clusters at their solution scores ($Solution_{Average}$ = 1.35; $Solution_{Total}$ = 5.30); however, they achieved their SSKT posttest scores (Posttest = 76.52) slightly above the mean score of School B (Posttest = 75.02), which is similar to Cluster 4. In terms of the solution submission rate, approximately 60% of the students in this group submitted at least one solution, and these students saved two aliens (i.e., 2.00) on average during the entire gameplay period. Interestingly, these students showed similar behavior patterns with Cluster 3 of School A regarding the amount of new and redundant information, which were relatively lower than the other

metrics in this group. This finding also supports that new information is positively related to the prediction of in-game performance.

Approximately 35% of the students are centered in Cluster 2 (n = 18, 35.29%), which is the biggest group in School B. Like the largest group in School A (i.e., Cluster 1), this group achieved the lowest after-game performance scores (Posttest = 69.08) and relatively lower in-game performance scores ($\text{Solution}_{\text{Average}}$ = 1.50; $\text{Solution}_{\text{Total}}$ = 7.61). The mean ranks of five inquiry behaviors were lower than those of the other three clusters. This group also shows the lowest solution submission rate; that is, about 44% of the students in this group did not submit any solution during the gameplay.

About 26% of the students in School B (n = 13, 25.49%) are in Cluster 3. This group achieved the highest after-game performance scores (Posttest = 81.94) and relatively higher in-game performance scores ($\text{Solution}_{\text{Average}}$ = 2.40; $\text{Solution}_{\text{Total}}$ = 11.38). The proportion of the students from Cluster 3 submitting at least one solution was approximately 77% (n = 10); these students saved between three and four aliens (i.e., 3.46) on average during the gameplay. This group launched six probes on average, which is close to the average number of launched probes in School B (i.e., 5.43). However, they received relatively higher new information (mean rank = 35.65) and lower redundant information (mean rank = 25.81), compared with the other groups. Notably, the number of repeated trials and errors of this group ranked the third with 17.69 and 20.62.

The rest of the students of School B (n = 10, 19.61%) are centered in Cluster 4. The radar plot shows the biggest pentagon, indicating the most active behaviors among all groups. The behavior patterns are similar to Cluster 2 of School A. The students in

both groups performed better at their in-game performance, but lower at the after-game performance than the other groups. Most students in this group (n = 8, 80%) submitted at least one solution during the gameplay; these students saved about 3.50 aliens on average.

Overall, the cluster analyses above suggest four unique behavior groups across two schools: (1) Lack of activity group, (2) Average activity group, (3) Trial-and-error group, and (4) Best performance group (see Table 19). To further interrogate the group membership and learning performance, two metrics were particularly selected: the amount of new information and the amount of redundant information, which are the significant predictors as found in the results of regression analyses. The researcher then presented multiple-layered information in a single view (see Figure 16) to illustrate the relationship between the amount of redundant information (Y-axis) and the amount of new information (X-axis) of four inquiry groups along with their SSKT posttest scores. Figure 16 shows four quadrants divided by two solid lines; that is, the horizontal solid line represents the average of amount of redundant information of all students, and the vertical solid line represents the average of amount of new information. Each point presents each inquiry group, and the size of the point indicates an average posttest score of each group as shown in Figure 16. In support of the regression analyses, the visualization in Figure 16 also showed the relationship between inquiry behaviors and performance scores and the characteristics of each inquiry group in terms of new information and redundant information.

Inquiry Behavior groups	School A	School B
Lack of activity group	Cluster 1	Cluster 2
Average activity group	Cluster 3	Cluster 1
Trial-and-error group	Cluster 2	Cluster 4
Best performance group	Cluster 4	Cluster 3

Table 19: Four Inquiry Behavior Groups of School A and School B

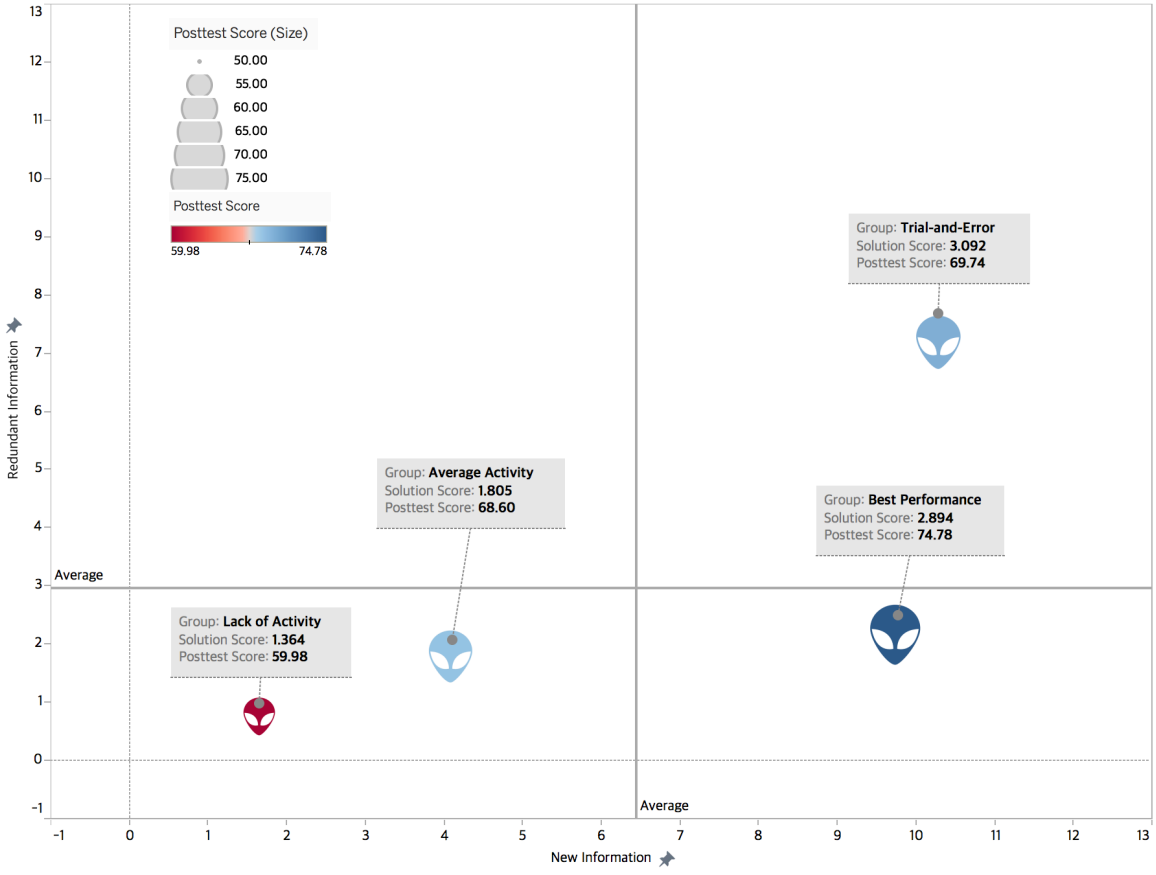


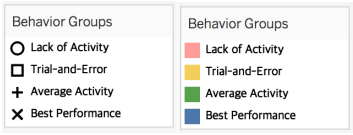
Figure 16: Redundant Information and New Information with Posttest Scores by Inquiry Behavior Groups

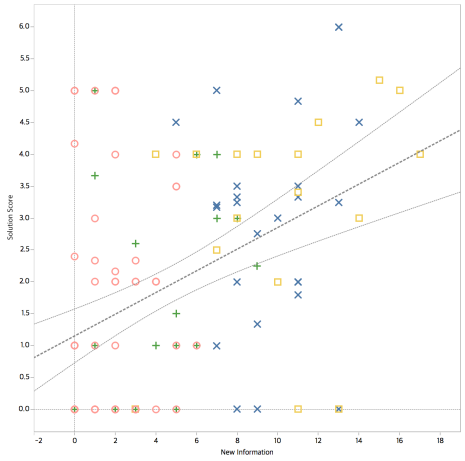
Note. See the interactive visualizations at http://tinyurl.com/utar-analytics

The two inquiry groups—relatively lower performing groups—are positioned at the bottom-left quadrant, indicating the students tended to receive both less amount of redundant information and new information. However, the average activity group received more new information and higher scores from their submitted solutions and SSKT posttests than the lack of activity group. Another group, trial-and-error group, is positioned at the top-right quadrant, indicating these students were prone to sending many probes and less likely reflect on the returned feedbacks from the probes. Last group, the best performance group, positioned at the bottom-right quadrant received relatively greater amount of new information compared with their amount of redundant information. Compared with the trial-and-error group, this group received slightly less amount of new information, indicating the students in the best performance group seemed to find a solution quickly using the less amount of information necessary.

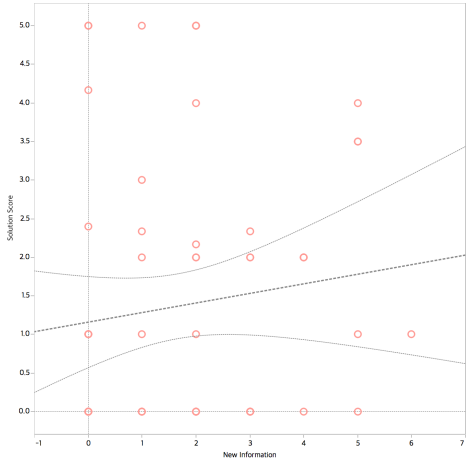
To further investigate the members in each group, the researcher visualized the relationships between the new information metric and in-game performance and the redundant information metric and after-game performance in a scatter plot (see Figure 17). In the previous analysis, the amount of new information was a significant predictor of in-game performance, and the amount of redundant information was a significant predictor of after-game performance. Each visualization includes the grey dotted trend line with the confidence band. While Figure 17 (b) reveals non-linear relationship between the amount of redundant information and posttest scores among all students, Figures 17 (d) and (f) show different trends between the variables when specifying each cluster. For instance, the relationship between the amount of redundant information and

posttest scores of the best performance group shows negatively stronger than the relationship of the lack of activity group. Figures 17 (c) and (e) show the students who received less amount of new information (i.e., between 0 and 6) were assigned into the lack of activity group, while the students who received greater amount of new information (i.e., between 5 and 14) were merged into the best performance group.

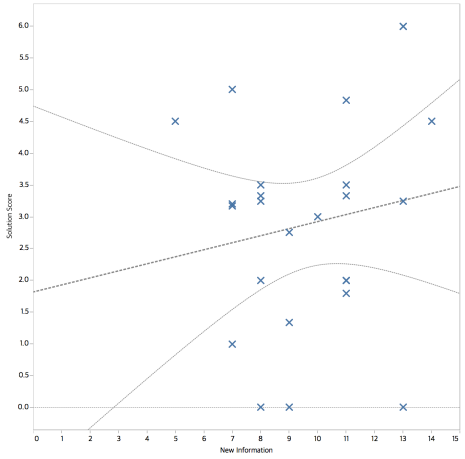




(a) New Information and Solution Scores of All students

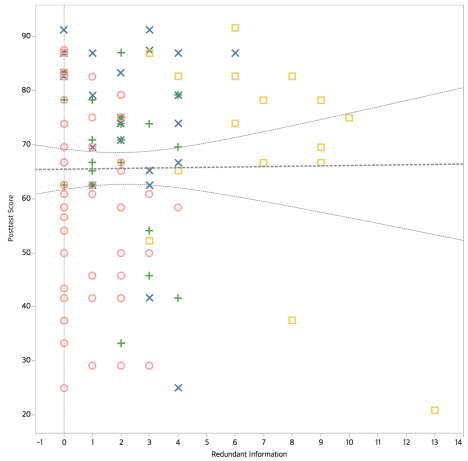


(c) New Information and Solution Scores of Lack of Activity Group

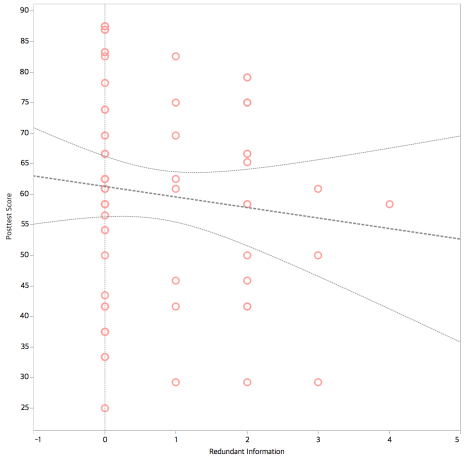


(e) New Information and Solution Scores of Best Performance Group

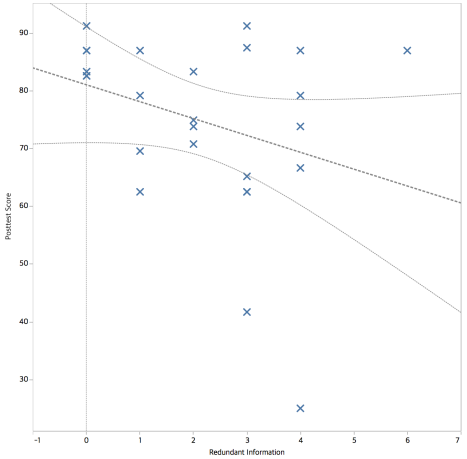
Figure 17 (continued)



(b) Redundant Information and Posttest Scores of All students



(d) Redundant Information and Posttest Scores of Lack of Activity Group



(f) Redundant Information and Posttest Scores of Best Performance Group

Figure 17: Redundant Information and New Information with In-game and After-game performances by Inquiry Behavior Groups

Note. See the interactive visualizations at http://tinyurl.com/utar-analytics

Summary of Analyses on Effect on Science Knowledge

The purpose of the third and fourth research questions regarding the effect on science knowledge is first to address the relationship between students' learning performance (i.e., in-game and after-game performances) and their inquiry behaviors, which emerged as students engaged with Probe Design Center in the serious game. The second purpose is to identify any unique behavior groups using students' scientific inquiry behavior metrics.

To address the third research question, two hierarchical regression analyses were carried out to investigate how much extra variation in students' learning performance (i.e., average solution scores as the in-game performance and SSKT posttest scores as the after-game performance) can be explained by the addition of scientific inquiry behavior variables generated from Probe Design Center. The predictors are the four behavior variables (i.e., number of repeated trials, amount of new information, amount of redundant information, and number of errors), and the two covariates are the number of launched probes and the school classification (School A, School B). In order to increase the sample size and control for a school, an additional sample of 51 sixth graders from School B was included to understand students' scientific inquiry behaviors in Probe Design Center across schools and build a model controlling for a school. The hierarchical regression analyses confirmed that the addition of scientific inquiry behavior variables to the predictions of in-game performance (i.e., solution scores; $R^2 = .188$, $F(4, 126) = 4.859$, $p < .01$, adjusted $R^2 = .149$) and after-game performance (i.e., SSKT posttest scores; $R^2 = .299$, $F(4, 126) = 8.942$, $p < .001$, adjusted $R^2 = .265$) led to statistically significant increases.

In addition to regression analyses, the researcher further conducted *k*-medoids cluster analyses with five game metrics to discover any distinctive inquiry behavior groups in each school and then examined the characteristics of each group. The results confirmed four unique inquiry behavior groups across two schools. The first group showed that the mean ranks of all five inquiry behaviors were lower than those of the other three groups; that is, this group can be categorized as a lack of activity group. This group performed relatively lower at both in-game and after-game performances. The second group overall showed an average activity and achieved relatively lower in-game performance scores and higher after-game performance scores. The students received the amount of new information below or slightly above the average of two schools and the amount of redundant information below the average. The third group showed the highest activity and appeared to possess a trial-and-error approach. They performed better at their in-game performance, but worse at their after-game performance. Compared with the other groups, they launched the most number of probes and collected a significant amount of new and redundant information. The last group appeared to be the best performance group. These students received the highest amount of new information with less redundant information, errors, and repeated trials, compared with the other groups.

VISUALIZATION FOR JUST-IN-TIME SUPPORT

The last purpose of this study is to address the challenge of understanding students' problem-solving processes through information interpretation derived from the large amount of user-generated data in the serious game, *Alien Rescue.* The researcher focused on Solar System Database and Probe Design Center and visualized students' in-game activities to support teachers to monitor students' problem-solving processes

through the visualizations. The following research question was asked to address the challenge of understanding students' problem-solving processes in *Alien Rescue*:

5) How can visualizations help to illustrate data-driven evidences of students' in-game behaviors to provide teachers just-in-time support?

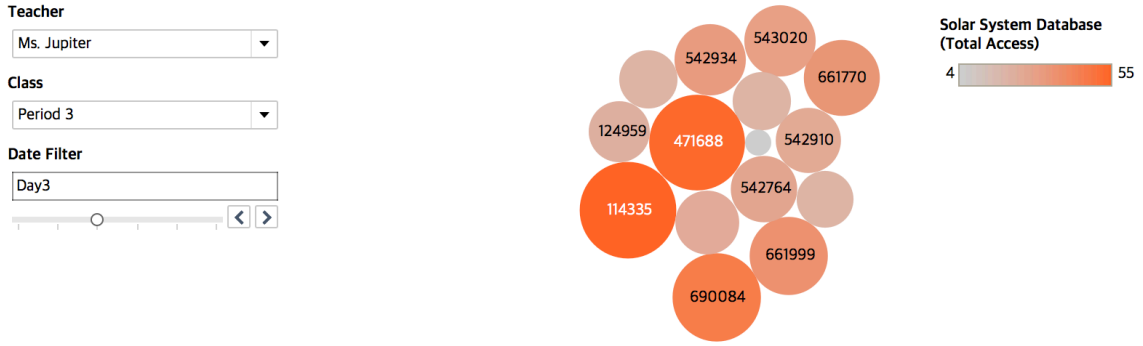
Research Question 5: How can visualizations help to illustrate data-driven evidences of students' in-game behaviors to provide teachers just-in-time support?

In the previous studies using this game in the classroom, one noticeable finding was that teachers needed some support to provide just-in-time feedback regarding how students were using this game. Therefore, one of the research questions in this study is if visualizations could be used to provide teachers just-in-time feedback based on data-driven evidence. To address this question, the researcher conducted the classroom observation, in which the teachers provided the students a project checklist presenting each step of problem-solving processes: (1) solar system research, (2) alien species research, (3) elimination chart, (4) probe prototype, (5) probe design, (6) probe launch, (7) probe results, and (8) recommendation. The students mainly filled out a paper-based worksheet for each step. During the gameplay, the teachers emphasized their worksheets must be approved by the teacher before they proceed to next step. The teachers obviously tried to monitor their students' progress by checking the worksheets and guide their problem-solving processes. However, the classroom observations revealed many students attempted to access different tools that they were not supposed to use in the step based on the teachers' checklist. Interestingly, many students were still in the solar system research part almost until the last day of gameplay. Although the teachers attempted to check out the worksheets of the less productive students, it seemed hard for the teachers to facilitate the students' problem-solving processes within the limited time for each class period.

Based on the classroom observations, the main challenges in the classroom use of *Alien Rescue* were that many students lacked the requisite knowledge of the productive in-game tool use within this open-ended serious game environment and required the significant teachers' support. These challenges therefore highlighted the teachers' advanced needs of tracking their students' problem-solving processes, instead of visiting an individual student's desk to check their worksheets. Since *Alien Rescue* is designed as a unit in the science curriculum, it aligns with the National Science Education Standards and Texas Essential Knowledge and Skills (TEKS). Particularly, Solar System Database contains information about our solar system including our sun, the nine planets, and ten of the moons, which is directly related to the science curriculum in school. The classroom observation also revealed that Solar System Database is one major in-game tool that the teachers wanted to track their students' research progress of and many students were struggling with. Beyond that, the previous analyses suggested the game metrics within Probe Design Center can be an indicator of students' in-game or after-game learning performance. This study examined the students' usage of Solar System Database and Probe Design Center and visualized their activities using diverse techniques in *Tableau*.

The findings from this study could inform the development of dashboard for teachers. Figure 18 shows the example of teachers' interactive dashboard to provide students' Solar System Database usage, which enables teachers to track daily Solar System Database accesses of students in each class. The bubble chart on the top right corner (Figure 18 (a)) displays students' aggregated accesses to Solar System Database up to the selected date by a date filter in cluster of circles format. The individual circle indicates each student, and both the size and color depth of each circle show the total access of Solar System Database by an individual student; that is, the bigger (or the darker) the circle, the more access the student made. Figure 18 (b) displays the list of

individual students' accesses to each planet system (i.e., planet and its moons) in a highlight table format. The number with color depth in each cell indicates the total number of access (i.e., the darker the violet, the more access the student made). The information in this table can be sorted by any planet system. Additionally, the interactive feature of *Tableau* was applied to these visualizations; that is, the bubble chart can be used as a filter. For example, a teacher can display only one student's Solar System Database usage by clicking a student who showed less access by the selected date in the bubble chart and then review which planet systems the student has not yet accessed in the highlight table.



(a) Filters (Left) and Bubble Chart of Students' Solar System Database Access (Right)

	Sun	Mercury	Venus	Earth System	Mars System	Jupiter System	Saturn System	Uranus	Neptune System	Pluto System
110989	0.000	0.000	0.000	0.000	0.667	0.000	0.500	0.000	0.000	0.500
124959	0.000	2.000	2.000	0.500	2.667	1.400	0.500	1.000	0.000	0.000
471688	0.000	4.000	4.000	1.500	1.667		2.000	2.000	3.500	2.000
542764	0.000	1.000	4.000	0.500	1.000	1.800	2.000	2.000	0.500	0.500
542871	0.000	1.000	1.000	0.000	0.667	1.400	2.000	3.000	0.500	0.500
542910	0.000	1.000	0.000	0.000	0.667	2.800	2.000	1.000	1.000	0.500
542934	0.000	4.000	2.000	0.500	1.333	2.000	2.000	3.000	1.000	0.500
543020	0.000	1.000	0.000	0.500	1.667		1.000	1.000	1.000	0.500
662025	0.000	0.000	0.000	0.000	0.000		1.500	1.000	0.500	0.500
402481	1.000	4.000	3.000	1.000	0.667	1.000	1.000	1.000	0.000	0.000
661872	1.000	2.000	5.000	0.500	1.333	1.200	0.000	0.000	0.000	0.000
114335	2.000	5.000	9.000	0.500	3.000		2.500	6.000	0.500	0.000
661770	2.000	3.000	2.000	0.000	1.333		2.000	2.000	0.000	0.500
661999	2.000	4.000	8.000	0.500	0.667	1.600	2.000	7.000	0.000	0.000
690084	3.000	5.000	7.000	1.000	3.667	2.200	2.000	3.000	0.000	0.000

(b) Highlight Table of Students' Solar System Database Access

Figure 18: Solar System Database Usage of Each Class

Note. A number in each circle (a) and the first column (b) indicates the school ID of each student. See the interactive visualizations at http://tinyurl.com/utar-analytics.

In addition, a teacher can track more in-depth Solar System Database usage of an individual student as shown in Figure 19. Figure 19 applied the background image of our solar system to visualize how many times an individual student has accessed each planet and moon up to a selected day. The color of a map marker indicates each planet system (e.g., a yellow marker indicates Jupiter and its moons), and the size of the marker shows

an accumulated number of visits to each planet or moon. Two filters on the left side can be used to select an individual student and a date of gameplay.

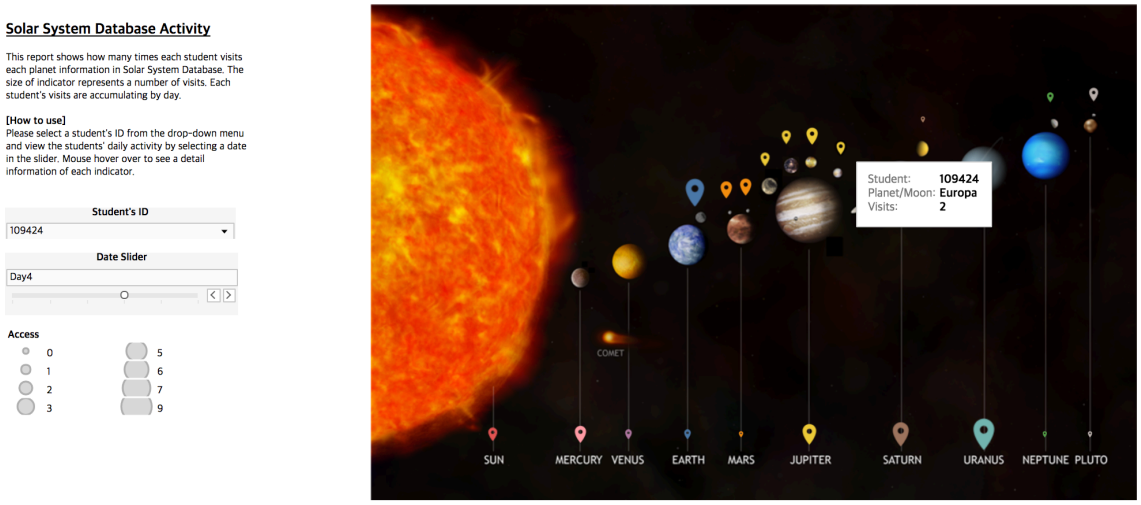


Figure 19: Solar System Database Usage of Individual Student

Note. Two filters on the left side are a Student's ID filter and a Date filter. A color of map marker indicates each planet system (e.g., a yellow color for Jupiter and its four moons). See the interactive visualizations at http://tinyurl.com/utar-analytics

The analyses conducted in the previous research questions suggested two significant game metrics generated as students interacted with Probe Design Center—the amount of new information and the amount of redundant information. Probe Design Center provides students authentic scientific inquiry learning experience, in which they can practice generating, testing, and evaluating a hypothesis by designing a probe with authentic space exploration technology. Throughout this experience, students can figure out that their decisions will impact the data—returned from the designed probe—which will challenge them to learn from their mistakes. The classroom observations revealed that many students repeatedly designed a probe with the same condition (e.g., same instruments or same probe type) with their previous probes, or launched several probes to

the same destination. The students struggled with errors returned from the launched probes and requested support from their teacher. However, the teachers often got distracted by some misbehaviors of students (e.g., chatting loud with other classmates, listening music, surfing internet). One teacher seemed frustrated when some students progressed further and requested any guidance from her in their probe design research part, since she spent most of time to arrange students who could not concentrate on their work to a separate seat from their group table.

To facilitate students' scientific inquiry process and address their challenges in this game, a teacher needs to grasp the idea of an individual student's interaction with Probe Design Center. The results of regression analyses particularly showed that the amount of new information can be a positive indicator of students' in-game performance, while the amount of redundant information can be a negative indicator of students' after-game performance. Figure 20 shows the amount of new information and redundant information of students in each class. Since the amount of new information is a positive predictor of learning performance, the orange color indicates that a student received less amount of new information compared with the school average (grey dotted line), while the green color indicates more than the average. As the amount of redundant information is the negative predictor, the red color shows that a student received more amount of redundant information compared with the school average (grey dotted line), while the blue color shows less than the average. The reference line (grey dotted line) supports teachers to figure out to what extent the amount of game metric of each individual student differs from the school average.

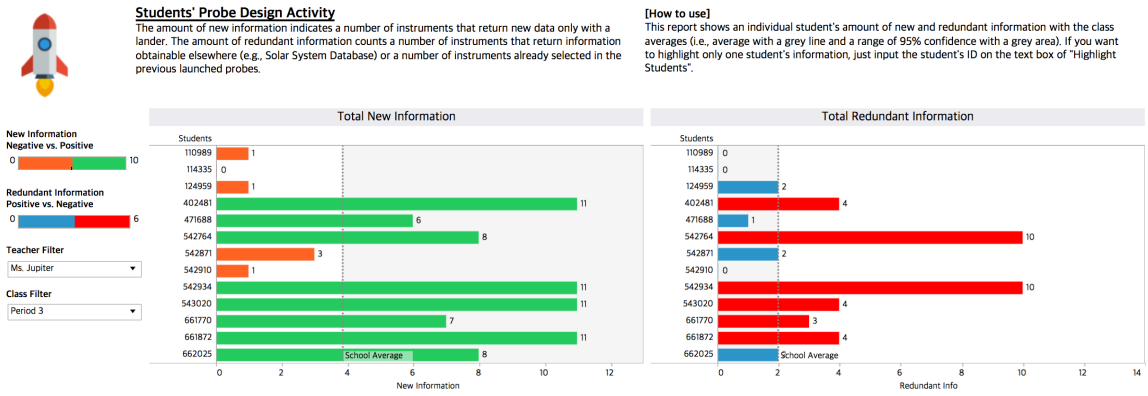


Figure 20: Probe Design Activity of Individual Students in Each Class

Note. See the interactive visualizations at http://tinyurl.com/utar-analytics

Summary of Analysis on Visualization

The purpose of the research question is to address if visualizations based on data-driven evidences of students' problem-solving processes can provide teachers just-in-time support. The classroom observations and results of previous analyses suggest two in-game tools which are critical to students' problem-solving processes: Solar System Database and Probe Design Center. Especially, Probe Design Center supports students' authentic scientific inquiry process by allowing them to discover that their choices will impact the data they receive. But, at the same time, this tool challenges them to learn from their mistakes and operate strategically. The researcher therefore proposed interactive visualizations to inform teachers not only an individual student's diverse tool usage, but also the student's progress or challenge in their problem-solving process involving in Solar System Database and Probe Design Center.

Chapter 5: Discussion

Summary of Research Findings

This study seeks to investigate sixth-grade students' scientific thinking processes in a three-week space science unit with an open-ended serious game, *Alien Rescue*, through using statistical methods in combination with data mining and visualization techniques. First, this study intends to identify navigation behavior patterns—as captured by the students' gameplay data—between non-at-risk and at-risk students. Second, this study seeks to examine the relationship between students' scientific inquiry behaviors emerged as students engaged with Probe Design Center in this serious game and their learning performance. Lastly, the study intends to address if visualizations based on data-driven evidences of students' problem-solving processes can provide teachers just-in-time support.

First, this study examined what differences exist between at-risk and non-at-risk students' navigation behavior patterns. The findings showed the non-at-risk group had significantly higher improvement on science knowledge than the at-risk group. The researcher further examined whether the daily frequencies of each in-game tool use differed between at-risk and non-at-risk groups, and found the non-at-risk students tended to access more often overall in-game tools than the at-risk students during the later days. The researcher integrated LSA and cSPADE to identify two groups' navigation behavior patterns in this game context. The results from both methods revealed that problem-solving strategies were differently used between the non-at-risk and at-risk students within this environment through the six days of their gameplay period.

Second, the researcher investigated the relationship between students' inquiry behaviors (i.e., game metrics) emerged as students engaged with Probe Design Center in

the serious game and their learning performance (in-game and after-game performances). The classroom observation and the gameplay data revealed that only 84 students (42.85%) accessed Probe Design Center; therefore, the researcher included additional gameplay data from another school to build a prediction model controlling for a school. Due to the different total days of gameplay, one game metrics, the total number of launched probes, was also treated as a covariate. The results from two hierarchical multiple regressions showed that the addition of four game metrics (i.e., amount of new information, amount of redundant information, number of errors, and number of repeated trials) improved the prediction of in-game and after-game performance over and above school and number of launched probes alone. The researcher further conducted cluster analyses with five game metrics to discover any distinctive inquiry behavior groups in each school and examined the characteristics of each group. The results confirmed four unique inquiry behavior groups across two schools: (1) Lack of activity group, (2) Average activity group, (3) Trial-and-error group, and (4) Best performance group.

Lastly, the researcher addressed if visualizations based on data-driven evidences of students' problem-solving processes can provide teachers just-in-time support. The classroom observations and results of previous analyses suggest two in-game tools that are critical to students' problem-solving processes: Solar System Database and Probe Design Center. Therefore, this study proposed interactive visualizations to inform teachers not only of an individual student's diverse tool usage, but also of the student's progress or challenge in their problem-solving process.

In the following section, the results are further discussed in the context of the literature. The directions of future research are suggested along with the limitations of this study.

IDENTIFYING NAVIGATION BEHAVIOR PATTERNS

Serious games have the potential to develop students' scientific thinking skills, in which the difference between experts and novices exists during their problem-solving processes (Dreyfus & Dreyfus, 2005; Jonassen, 2000; van Merriënboer, 2013; Wiley, 1998). Research has identified the challenges of understanding students' behaviors in complex game systems (Gobert et al., 2015; Hmelo-Silver & Azevedo, 2006). Specifically, students with poor academic performance, a lack of motivation or self-direction—that is, characterized as at-risk students (Hammons-Bryner, 1995)—may face the challenges during the processes of problem-solving in the serious games. However, there is little known about the at-risk students' learning behaviors in the serious games. This study identified the latent learning behavior patterns of at-risk and non-at-risk students to understand their experience with the serious game.

In the analysis of the differences in the posttest scores after accounting for the pretest scores between at-risk and non-at-risk groups, the non-at-risk group exhibited greater learning gains after the gameplay ($F(1, 193) = 16.911$, $p < .01$). Further investigation of the at-risk and non-at-risk students' navigation patterns as they interacted with various in-game tools revealed that the learning gains on science knowledge can be associated with different navigation behaviors between the two groups.

The examination of a daily frequency of each in-game tool showed no significant difference between the two groups' tool frequencies in the early days, but notable differences on the later days, in which the non-at-risk group increasingly used multiple in-game tools (i.e., Alien Database, Probe Design Center, Mission Control Center, and Communication Center). The previous research found the students with high performance showed increasing use of these tools from early stages to later stages of the problem-solving process (Liu et al., 2016). Additionally, the increasing use of multiple tools from

early stages to later stages in problem-solving appeared to be an indication of the improvement of students' tool use strategies (Liu & Bera, 2005). The finding of this study, therefore, suggests the non-at-risk group exhibited more productive use of the tools.

To discover in-depth navigation behavior patterns of each group, this study applied the integrated method of sequential pattern analyses using cSPADE algorithm with LSA. Clark et al. (2012) noted that the main challenge of sequential pattern mining is the inability to provide the information of the exact time each sequential pattern occurs. Therefore, the integrated method was conducted with the navigation data of each day to identify daily sequential patterns. The results revealed that the two groups (i.e., non-at-risk and at-risk groups) exhibited different problem-solving behaviors as they interacted with different tools at different times throughout the entire gameplay.

During the first two days, both non-at-risk and at-risk students tended to navigate the environment by exploring as many in-game tools as possible and switching between different tools. This navigation pattern indicates that the students attempted to understand what the function is of each tool and how different tools can be used together on their first day of problem-solving. On the second day, the at-risk students were more prone to experiencing the tools which are more appropriate to use during the later days of problem-solving processes (i.e., Probe Design Center and Mission Control Center). This finding is consistent with previous studies; low-performing students tended to use more fun tools (Kang et al., 2017; Liu et al., 2015). Since the second day is still the early stage of problem-solving, students are expected to access mainly Solar System Database and Alien Database to get an overview of the problem and gather information about planets and alien species (Liu et al., 2009, 2015).

On the later days, the at-risk students did not show any significant transition between Alien Database and any other tools, but the transition from Alien Database to Alien Database. In addition, most of the navigation patterns of at-risk students were sequences where one tool follows the same tool after Day 3. This type of patterns indicates the at-risk students repeatedly visited the same tools and did not make any further progress of integrating important information provided across different tools. Research shows expertise critically influences students' problem-solving processes (Dreyfus & Dreyfus, 2005; van Merriënboer, 2013). The previous studies found that low-performing students seemed to struggle to find out important information embedded in different tools (Kang et al., 2017; Liu et al., 2009, 2015, 2016). Another study on the experience of at-risk students with *Alien Rescue* noted a majority of the at-risk students showed boredom due to the confusion or difficulty dealing with a less guided instruction (Samsonov et al., 2006). Similarly, during the classroom observations in this present study, the at-risk students were overwhelmed by the amount of information provided in Solar System Database and stagnated in the step of solar system research. These students seldom reflected on their problem-solving processes, but merely asked simple questions to their teacher. That is, most of the students ended up with the solar system research step at the end of the gameplay and therefore, could not move forward to the next step of probe design activity. Thus, the navigation behaviors of at-risk students, which had remained stagnant on the later days, could be possibly explained by their lack of structured knowledge and metacognitive skills in their problem-solving processes, which suggests the needs of support for the at-risk students. Research highlights novice students possess a low level of meta-awareness regarding the strategic tool use in different cognitive processes (Simons & Klien, 2007). Novice students need support for

developing a procedural model such as developing plans for solving a problem in the game environment (Bogard et al., 2013).

Together with the U-test results, the significant patterns discovered by cSPADE and LSA suggest the most critical sequences that support students to progress in solving a complex problem in this game. For example, the transition from Mission Control Center to Communication Center was significant only for the later days in the non-at-risk group, which indicates these students proceeded to make a recommendation for each alien based on the critical information gathered from Mission Control Center. During the same period, the at-risk group repeatedly accessed the same tools (i.e., mcontrol → mcontrol; communication → communication). The classroom observations also revealed how students became more strategic in tool use on the later days. As an example, one of the non-at-risk students tended to use his own prior knowledge to find out a possible home for the alien species, Sylcari, which turned out to be inappropriate home choices. During the conversation with other classmates, this student mentioned one documentary about Ganymede from the National Geographic channel and believed Ganymede as a possible home for Sylcari. He was first frustrated about his failure. However, he soon figured out his prior knowledge was not correct and needed to gather the additional temperature data in Mission Control Center, which was not provided in Solar System Database. He accessed Mission Control Center to check out the returned data from the probe he launched in Probe Design Center. Finally, he eliminated Ganymede from the list of possible homes and opened Communication Center to write a correct solution. Advanced prior knowledge of the domain forwards cognitive processes or knowledge construction processes (Livingston & Borko, 1989). Previous research (Bogard et al., 2013) noted that students with prior domain knowledge had a tendency to misrepresent the problem and ignore the functions of cognitive tools that allow them to collect contextual knowledge

about the aliens, habitats, and other relevant factors. The various in-game tools enable students to facilitate their information processing by coordinating multiple cognitive skills in this game (Liu & Bera, 2005). The non-at-risk group tended to explore the available tools and discover their capabilities at the beginning of gameplay and then develop their own strategies for how and when to effectively use the tools. The results of ANCOVA demonstrated the non-at-risk students performed better at the posttest SSKT scores (i.e., after-game performance). That is, the non-at-risk group seemed to achieve both conceptual and procedural knowledge while solving a complex problem throughout problem-solving processes.

This study aggregated students' navigation data by using the student-based sequence modeling. Therefore, the identified frequent sequential patterns across a group of students can indicate frequent or common learning behaviors within this group. The integrated method of two different sequence analyses confirmed different navigation patterns between the at-risk and non-at-risk students are associated with their learning gains on science knowledge, which is consistent with the findings of previous studies (e.g., Kang et al., 2017; Liu et al., 2016). The previous studies revealed different problem-solving strategies between low- and high-performing students. For example, less productive tool use can affect their learning performance, as all students in a class were given the same amount of time to solve a problem in the game. Accordingly, this study found that at-risk students possess the behavior patterns similar to low-performing students, while non-at-risk students exhibit the productive strategies similar with the behavior patterns of high-performing students.

This study conducted two sequential analyses: lag sequential analysis (LSA) and sequential pattern mining with cSPADE algorithm. LSA examines each group's transitional probabilities of all the pairs of in-game tools and provides a significant

sequence, which deviated from their expected values, while cSPADE revealed the navigation behaviors across each group of students (*min_sup* > 0.3). That is, more than 30% of students within the group used the constrained frequent sequences. Although cSPADE can identify a frequent sequence of a number of items, the results showed only the sequences of three items at most. This can be explained by the fact students do not typically make the large number of "open" actions in this game context due to the limited amount of time per day (i.e., approximately thirty to forty minutes) spent in actually using the game (Kang et al., 2017; Liu et al., 2016). Another reason can be due to the two-layer interface structure: the first layer consisting of four main tools and the second layer consisting of six tools, in which the tools in the first layer cannot be overlaid together. Also, the purpose of this study is to examine the daily sequences of tool use in a group of students, which decreased the data size; therefore, a sequence with many items was not expected.

Understanding the sequential structure of interaction in serious games with science context supports in-depth understanding of learners' problem-solving processes (Chung & Baker, 2003; Hou, 2015; Pohl et al., 2016). The integrated method of LSA—as a statistical approach—and cSPADE—as a data mining approach—helped to explore students' various navigation behavior patterns and determine the different problem-solving processes between the at-risk and non-at-risk students. The results of LSA provided detailed daily navigation behaviors between the two groups based on transitional probabilities of all possible pairs of in-game tools. Although the findings of cSPADE revealed only a few sequential patterns when most of the students mainly used Solar System Database (i.e., Days 2-3), the overall results supported the group differences found in the LSA results. Furthermore, the state transition diagrams based on

adjusted residuals (see Figures 6-11) allow in-depth understanding of the differences of the two groups' significant sequences.

EFFECT ON SCIENCE KNOWLEDGE

To discover significant game metrics to predict students' in-game and after-game performances (i.e., average solution scores as the in-game performance and SSKT posttest scores as the after-game performance) after controlling for a school, an additional sample of 51 sixth graders from School B was included. Two hierarchical regression analyses were conducted to investigate how much extra variation in students' learning performance can be explained by the addition of scientific inquiry behavior variables generated from Probe Design Center. The findings confirmed the addition of scientific inquiry behavior variables to the prediction of both in-game and after-game learning performance (i.e., average solution scores and SSKT posttest scores) led to statistically significant increases. In addition to the regression analyses, this study further conducted a k-medoids clustering analysis with *pam* algorithm using five game metrics of each school and found four clusters in each school. Then, the researcher cross-examined the identified clusters in each school and discovered four distinctive inquiry behavior groups across the two schools: (1) Lack of activity group, (2) Average activity group, (3) Trial-and-error group, and (4) Best performance group (see Figure 16).

The first cluster group, Lack of activity group, showed overall a lack of activity, indicating they did not actively access Probe Design Center and achieved the lowest after-game performance. This group performed relatively lower at both in-game and after-game performances. Probe Design Center provides additional information about planets that would not be otherwise possible to find anywhere else. Previous studies

showed Probe Design Center supports students' hypothesis testing, in which students can collect critical information that they need to eliminate or confirm potential home choices and finalize a solution of each alien (Kang et al., 2017; Liu et al., 2009). Therefore, the characteristics of a lack of access to Probe Design Center suggest that the students had a tendency to misuse tools with the cognitive processes or use ineffective tools for their knowledge development (Bogard et al., 2013). Ultimately, the students in this group were unsuccessful performing both in-game and after-game performances due to the limited evidences—gathered from Probe Design Center—to examine the affordances and constraints of a potential home for the aliens.

The second behavior group, Average activity group, overall showed average activity. Specifically, the students showed most inquiry behaviors (errors, repeated trials, and launched probes) slightly above the average of the two schools. The clusters from both schools showed the students performed relatively lower at their in-game performance, but higher at the after-game performance (i.e., the second highest SSKT posttest scores). The regression analyses found the amount of new information was positively related to students' in-game performance. In terms of the amount of new information, the School A's was close to the school average, and the School B's was lower than their average. Therefore, these students' low in-game performance can be explained by the fact that both clusters showed the students received the amount of new information far below the school average. By applying new information students additionally obtained from the launched probes, students were more likely to find a solution, since the obtained information can be the critical clues to find a possible home for the aliens. On the other hand, the students in both clusters showed the amount of redundant information to be below the average. This finding confirmed the results of regression analyses that the students with less redundant information performed better at

their after-game performance. The amount of redundant information counts the number of instruments that return information obtainable in Solar System Database or the number of instruments already selected in any previous launched probes. That is, the students in this group strategized using Probe Design Tool to build a probe to find out the missing information that they could not find elsewhere in the game environment. Therefore, students who received less redundant information demonstrated that they were more strategic in their problem-solving processes in this game context. Previous study (Bogard et al., 2013) noted the highly self-regulated students often showed evaluating outcomes, responding productively to their failures, and adjusting plans and strategies. Consequently, these traits helped the students to become successful problem solvers.

The third behavior group, Trial-and-error group, showed the highest activity among all groups. The tendency of this high activity suggests that the students appeared to possess a trial-and-error approach. That is, these students attempted many trials (launched many probes), evaluated their failures (errors or redundant information), and manipulated their hypotheses (designed new probe). However, the highest frequencies of the amount of redundant information and the number of repeated trials indicate the students did not successfully remedy their mistakes by interpreting errors or redundant information from previous probes since they also showed many repeated trials or redundant information. Students' hypothesis testing strategies in science education games are strongly associated with their learning effectiveness (Spires, Rowe, Mott, & Lester, 2011). The solution scores of these students demonstrated their higher levels of in-game performance. For example, they found plenty of new information that was the direct clues to find right homes for the aliens. This finding showed they had a tendency to send a probe without consideration of costs or feasibility and merely to seek information instead of testing their hypotheses. In addition, they tended to repeat incorrect manipulations by

launching probes to the same destinations with any previous probes. This tendency may make them to waste too much time to find out effective problem-solving strategies and to exacerbate backward reasoning processes (Chi & Bassock, 1991; Glaser, 1989; Thorndike, 1913). This ineffective reflective process ultimately did not yield learning gains in science content knowledge that was assessed by the posttest.

The last group, Best performance group, showed a spike in the amount of new information. The students showed the number of launched probes to be slightly above the average of the two schools. They received the highest amount of new information and relatively lower amount of redundant information compared with the other groups. Furthermore, they obtained additional information—required to eliminate or confirm their solutions—with less errors or repeated trials. Gick (1986) proposed the model of the problem-solving process, in which learners who have previously solved a similar problem or have a high level of expertise can be more efficient to discover a solution scheme to solve a problem by avoiding any redundant iteration processes. Accordingly, the findings in this study suggest the students in this group may possess more effective problem-solving strategies—that is, explicit knowledge of how to design appropriate probes. Therefore, they could reduce time on their cognitive processes in this game. Students who showed this effective problem-solving strategies achieved the highest SSKT posttest scores, which demonstrates they gained greater science content knowledge throughout the entire problem-solving processes during the gameplay time—equally given to all students.

In sum, the regression models of five scientific inquiry behaviors and school variable to predict in-game and after-game performances were statistically significant. Specifically, the amount of new information appeared to be associated with students' ability to apply the information to complete a task within the game, and therefore

positively associated with students' in-game performance. The amount of redundant information was negatively associated with students' problem-solving strategies; that is, students who received less redundant information demonstrated effective problem-solving strategies in this game context and therefore performed better at after-game performance.

Novices appear to have difficulties to recognize the problem, find a workable solution path, retrieve the knowledge that is relevant to a particular task, or integrate meaningful information (Bransford et al., 2000). Recognizing the tools to best support developing solution procedures or building the associations between tools is not automatic behavior for novice learners. Tools that can offer guidance in developing their structural knowledge are suggested for students in the lack of activity group. Analyzing the problem and identifying the function of cognitive tools are critical to develop a procedural knowledge. The findings of this study are aligned with the previous study to examine students' self-regulation (Bogard et al., 2013), which suggest that students who possess a low level of self-regulation need support to readjust a plan and strategies after evaluating outcomes. Such needs can be embedded in the game environment as a new tool. When evaluating outcomes, students with the traits of the trial-and-error cluster may need scaffolds to recognize the gaps in their knowledge development and reflect their procedural knowledge to devise next steps using tools with a prompt to trigger their cognitive processes.

VISUALIZATION FOR JUST-IN-TIME SUPPORT

Visualization supports various purposes of different stakeholders: providing just-in-time feedback by tracking students' progress for teachers, monitoring learning

progress for self-reflection for students, or assessing game design or pedagogical effectiveness for researchers or game designers (Wallner & Kriglstein, 2015). In this study, visualizations were applied to address two purposes; First, to support the results of analyses and second, to see if visualizations can provide teachers just-in-time support based on data-driven evidences of students' problem-solving processes.

First, the visualizations in this study support the results of statistical analyses (i.e., ANCOVA, hierarchical regression analyses). The state transition diagrams based on adjusted residuals allowed in-depth understanding of the differences of significant sequences between at-risk and non-at-risk students (see Figures 6-11). The radar plots revealed the latent patterns of scientific inquiry behaviors between four groups—found as the results of cluster analyses—and confirmed that the groups' inquiry behavior patterns were distinctively different in this serious game (see Figures 14-15). The researcher also used two significant metrics (i.e., the amount of new information, the amount of redundant information) to further understand the cluster group membership (see Figure 16) and visualized the relationships between the two metrics and in-game and after-game performances in an interactive scatter plot (see Figure 17). The diverse graphical representation techniques using a juxtaposition strategy support better comprehension of differences among different student groups.

Second, as shown in Figures 18-20, the interactive visualizations in this study allow teachers to track students' problem-solving processes and get involved in their activities as needed. Furthermore, teachers can monitor students' activities at the level of the classroom and an individual student and therefore facilitate classroom management and assessment. Serious games with science context should allow learners to get involved in the problem-solving tasks by manipulating experiments and evaluating the results that enhance their problem-solving skills and motivation (Hou, 2015). The games can capture

what an individual learner is doing in the environments by tracking the student's learning process using the captured data to evaluate their inquiry process and measure their learning performance (Gobert et al., 2013; Quellmalz et al., 2009). Research on the computer-based environments reported the challenges of teachers to integrate inquiry into classrooms such as the lack of materials, technical support, or teacher preparation (Anderson, 2002; Welch, Klopfer, Aikenhead, & Robinson, 1981). This study not only showed the potential of visualization to facilitate the interpretation of the relationships among multiple data, but also provided empirical support for the use of diverse visualization techniques to support teachers' classroom use of the serious games.

CONCEPTUAL AND PROCEDURAL KNOWLEDGE IN SERIOUS GAMES

In serious games, learning occurs throughout the process of understanding the game system, conducting experiments, continuously adjusting problem-solving strategies, and communicating with other learners (Killi, 2007; Squire, 2008). Obviously, the findings of the current study showed that Probe Design Center is the major tool to allow students to experience interactive reflective processes by repeatedly testing their hypotheses in this game environment. Research on scientific problem-solving emphasized the importance of conceptual and procedural knowledge (Gott, Duggan, & Roberts, 2008; Klahr & Dunbar, 1988; Wiley, 1998). That is, scientific problem-solving is always involved in conceptual change, which can be achieved only when learners possess both conceptual and procedural knowledge. Specifically, conceptual change can occur when learners recognize that the previous conceptual understanding of scientific phenomena is wrong and therefore needs to be changed (Mayer, 2008). Bogard et al. (2013) showed that highly self-regulated students tended to evaluate outcomes and readjust their strategies to discover knowledge constraints and build dynamic mental

model of the problem in a serious game. Therefore, the researchers highlighted the role of self-regulation to solve a complex problem and the need of support for novices' knowledge development. These research findings are consistent with the characteristics of the best performance group identified in this study, indicating the students possess both conceptual and procedural knowledge throughout the gameplay.

On the other hand, many researchers have observed students' difficulties to understand complexities of system and scientific inquiry process (Gobert et al., 2015; Hmelo-Silver & Azevedo, 2006). Gobert et al. (2015) noted when students design experiments, they possibly collect limited evidence to test their own hypotheses, attempt only one trial or repeated trials with the same condition, or revise too many variables. This study also found that the two behavior groups—trial-and-error group and average activity groups—showed ineffective reflective processes and performed lower at the posttest scores than the other group with effective problem-solving strategies. Open-ended serious games intend to provide students an opportunity to continuously practice their decision-making and evaluation skills with multiple solution paths in a less guided learning environment (Liu & Bera, 2005; Spring & Pellegrino, 2011). However, the findings of this study suggest the characteristics of open-ended serious game can be overwhelming for some students; therefore, different students need timely adequate supports to develop their own problem-solving strategies.

NEEDS OF SERIOUS GAMES ANALYTICS

The inquiry in this study built upon previous research on understanding students' use of in-game cognitive tools and their cognitive processes in the serious game, *Alien Rescue*. Some studies were conducted by analyzing gameplay data (i.e., frequency or

duration of each tool use). For example, Liu and Bera (2004, 2005) investigated the use of cognitive tools across five contextual problem-solving stages (i.e. initial exploring, background research, hypothesis generation, hypothesis testing, and solution generation) through a cluster analysis of gameplay data (i.e., frequency of each tool use). The results showed that students were strategic in their tool usage over entire stages; that is, students used cognitive tools in more sophisticated ways during the later stages of problem-solving. The students' tool use was highly correlated with their performance. Liu et al. (2015, 2016) applied data visualization techniques to discover students' tool use patterns and identify any contributing factors to student variations. The results indicated different tool usage patterns between different groups of students; for example, high performing and mastery goal-oriented students tended to use the appropriate tools relative to each problem-solving stage.

A classroom observation or student interview is another type of data that were mainly used in previous studies. For instance, Liu et al. (2009) conducted a study with undergraduate students who played the serious game in a laboratory setting. Each student's activities in the environment were observed, and an interview was conducted to determine students' cognitive processes at a specific problem-solving stage. The authors confirmed the results from the previous studies of the strong association between cognitive tool use and cognitive processes. In another study with student interviews, Bogard et al. (2013) conducted the descriptive analysis (i.e., cross cluster analysis) using stimulated recall, think-aloud, and direct observation to address how students' application and frequency of cognitive processes and behaviors contributed to differences in performance outcomes and mental model development. The findings revealed that students with consistent self-regulation—the most expert-like learners—kept their cognitive processes in carrying out operations in each threshold of knowledge

development: 1) Building a procedural model, 2) building a structural model, 3) building an executive model, and 4) building arguments. The students developed their mental models through each threshold and thereby focused on the most relevant aspects of the problem solutions.

Overall, existing research on *Alien Rescue* provided empirical evidence that cognitive tools play a critical role to support students' problem-solving and activate their cognitive processes in this serious game context (e.g., Bogard et al., 2013; Liu & Bera, 2005; Liu et al., 2009, 2016). Research on serious games highlight the open-endedness of a serious game engages students' scientific problem-solving process, however, a game's complex system challenges researchers to understand students' diverse behavior (Squire, 2008). Especially, the use of traditional educational assessments is a great challenge in understanding how students learn complex skills through solving problems within open-ended serious game environments. However, extant research is mostly based on traditional assessments such as pre- and posttests or self-reported surveys, or general analytics metrics such as frequency and time-to-completion rate. Therefore, this present study indicates the importance of using gameplay data and game metrics to better understand students' behaviors and performances in different learning contexts.

The emergence of serious games analytics with growing opportunities of collecting massive gameplay data enables the tracking of sequences of actions during students' problem-solving as an evidence of learning performance in serious games and reduces claims of generalizability due to the data collected in context-specific situation. The current study conducted the integrated analysis of traditional statistical methods and data mining and visualization techniques using *in situ* gameplay data in order to discover meaningful patterns of students' cognitive processes and identify their diverse problem-solving strategies and the challenges students with different characteristics may face. In

addition, little is known about the potential meanings of a parameter of *in situ* data as a behavior indicator within this serious game context. Therefore, the researcher used game metrics—developed using gameplay data generated in Probe Design Center—as features of students' scientific inquiry behaviors and identified different inquiry behavior groups across different schools.

In this current study, the researcher integrated lag sequential analysis (LSA) and sequential pattern mining with cSPADE algorithm to identify latent navigation behavior patterns between the at-risk and non-at-risk students in this serious game. The findings discovered these two groups' significant sequential patterns of cognitive tool use on each day of gameplay and identified different problem-solving strategies between these two groups. This study also conducted a *k*-medoids clustering using *pam* algorithm to explore the potential cluster patterns of students' various inquiry behaviors in Probe Design Center across the school settings. The integrated method of cluster analysis and data visualizations enabled the investigation of in-depth scientific inquiry processes within a specific cognitive tool, Probe Design Center, and discover unique behavior groups across school settings.

IMPLICATIONS

Research on scientific problem-solving emphasizes that students need to possess both conceptual knowledge and procedural knowledge, which are the critical components of twenty-first century skills (Clark-Midura et al., 2011; Gott et al., 2008; Lederman et al., 2014; Wecker et al., 2013). The main challenge of open-ended serious games is to identify students' diverse solution paths in the complex game system (Squire, 2008). Identifying the sequential patterns of students' cognitive tool use in serious games can

provide insights of students' cognitive processes in the complex system of serious games. Prior research on *Alien Rescue* has identified students' cognitive processes in the game environment based on observation and interview data, and general game metrics such as frequency or duration of cognitive tool use (e.g., Bogard et al., 2013; Liu et al., 2005, 2009, 2016). The researcher in this study performed both LSA—as a statistical approach—and cSPADE—as a data mining approach—using the navigation data of non-at-risk and at-risk students each day. This integrated method enabled to identify in-depth students' cognitive processes in an open-ended serious game.

cSPADE can identify frequent sequences of a number of tools (Zaki, 2000, 2001). The results of this study showed only the sequences of three items at most, which can be explained by the fact that the students did not typically make the large number of "open" actions in this game due to the limited amount of time per day (i.e., approximately thirty to forty minutes) spent in actually using the game. This finding informs future studies of behavioral analysis in *Alien Rescue* can apply cSPADE, if there is a certain situation such as a large amount of gameplay data and more students' decision making allowed in a classroom. For example, a teacher does not provide any rigid guidance on students' problem-solving processes, rather facilitate them to find out their own problem-solving strategies by accessing various in-game tools during the entire gameplay, without any restriction. In addition, sequential patterns discovered by a sequential pattern mining are determined by the parameter, *support*, indicating the number of occurrences of the patterns. However, in some contexts, *support* may not always represent the significance of a pattern. For example, some researchers might be interested not in frequent patterns with many occurrences, but infrequent patterns with only few occurrences because these patterns are expected or surprising in a certain context (Esmaeili1 & Gabor, 2010). The researcher in this study thereby conducted another sequential analysis, LSA, as a

statistical technique to examine whether a sequence of tools achieves statistical significance among all sequential tool pairs in a certain group of students. This study conducted LSA to find out only significant two-tool sequences (i.e., *lag* 1) due to the lack of diverse tool use and limited amount of time spent in using the game of the students in the school (i.e., School A). LSA can be conducted to identify a sequence of more than two elements. However, researchers should determine a method after considering the factors such as a gameplay period, data size, and a teaching style to reduce time and cost during the analysis. In addition, Clark et al. (2012) noted one challenge of sequential pattern mining is that it cannot provide actual time information when sequences of actions occur. Therefore, this study conducted a separate sequential analysis for each day to better understand when or why students actually access certain in-game tools in identified sequences, which researchers should consider when conducting sequential analyses. The findings of sequential analyses in this study can be also applied to other future research of behavioral analysis dealing with understanding student cognitive processes in serious games.

Extant research on educational games is mostly based on traditional assessments such as pre- and posttests or self-reported surveys, or general analytics metrics such as frequency and duration of in-game tools. Therefore, discovering features as an indicator of diverse learning behaviors using gameplay data is essential to better understand students' behaviors and performances in the context of serious games. However, little is known about the potential meanings of a parameter of *in situ* data as a behavior indicator within the context of serious games. Therefore, this study contributed to developing game metrics generated in students' gameplay data from Probe Design Center as features of students' scientific inquiry behaviors and conducted a *k*-medoids clustering using *pam* algorithm. Using the game metrics, this research conducted the cluster analyses and

identified different inquiry behavior groups across different schools. The results of hierarchical regression analyses confirmed the addition of four game metrics (i.e., amount of new information, amount of redundant information, number of errors, and number of repeated trials) in Probe Design Center improved the predictions of both in-game and after-game performance over and above the covariates of school and number of launched probes alone. However, the full prediction model on after-game performance showed two covariates as significant predictors, indicating a school and the number of launched probes are significantly related to students' after-game performance even after the addition of game metrics, which is the limitation of this study. Therefore, this study suggests the need of considering a different school and the number of launched probes when building a prediction model. Another limitation is that the participants of School B only took the SSKT posttest, not the pretest; therefore, this study only considered the posttest scores as after-game learning performance to address the third and fourth research questions. In the future, researchers should consider the need of bigger data sets of gameplay data from diverse schools to understand students' scientific inquiry behaviors in Probe Design Center across schools and build a model controlling for a school. This will consequently reduce the claims of generalizability due to the data collected in context-specific situation. Future study is suggested to develop additional features relevant to students' scientific inquiry behavior to further understand their relationship with learning performance in this context of serious game.

Research on scientific problem-solving indicates that young students have difficulties in conducting scientific inquiry and a challenge of measuring younger students' inquiry strategies due to the difficulties to understand the complexities of system and scientific inquiry process (Gobert et al., 2015; Hmelo-Siver & Azevedo, 2006; Lederman et al., 2014). Therefore, game metrics that can measure students'

specific behaviors need to be developed to better understand how diverse students conduct scientific inquiry and facilitate their inquiry process in a complex learning environment such as *Alien Rescue*. Since scientific inquiry involves critical and logical thinking, traditional educational assessments do not demonstrate students' conceptual and procedural knowledge related to inquiry (Clarke-Midura et al., 2011; Gobert et al., 2013; NRC, 1996; Quellmalz et al., 2009). This research proposed the use of visualizations to support teachers to provide just-in-time guidance based on the game metrics identified in this study. Such information will help teachers to identify students' difficulties in their cognitive processes and why the students cannot immerse themselves in complex learning environments. Thus, teachers provide students a proper guidance to enhance their deeper reflection and increased learning.

Research highlights the potential of cognitive tools to support knowledge building and scaffold higher-order cognitive tasks within complex learning environments (Bogard et al., 2013; Jonassen, 2004). In this current study, the in-depth sequential patterns of cognitive tool use between the at-risk and non-at-risk students showed the diverse problem-solving strategies of these two groups. This finding highlights that students such as those placed at-risk need support for developing contextual and procedural knowledge for solving a problem in this game environment. Consequently, this study suggests additional tools that can provide guidance in developing students' knowledge. Such tool can prompt a window where students can monitor their knowledge development by answering a series of questions: what information they have found out so far, what information they still need to collect to find out a possible solution, and what tools they can use to gather the missing information. This current study can further inform practical guidelines for game developers to design different game levels and levels of guidance in

serious games with a complex problem for supporting students who possess a lack of skill in metacognition and self-regulation.

In sum, this research indicates the importance of using gameplay data and game metrics to better understand students' behaviors and performances in different learning contexts. Besides the benefits for teachers, the analyses in this study have provided the integrated method of traditional statistical analyses and data mining and visualization techniques of understanding in-depth students' cognitive processes in the context of serious games, which can be beneficial for game designers, researchers, and students.

CONCLUSION

Open-endedness of a serious game engages students' scientific problem-solving process. However, understanding how students learn complex skills through solving scientific problems is a challenge due to the complex learning systems. Recent research stresses the importance of using gameplay data to better understand individuals' learning behaviors and performances in the context of serious games. This study analyzed *in situ* data of the serious game, *Alien Rescue*, by applying the integrated method of traditional statistical analyses, data mining, and visualization techniques to identify in-game behavior patterns and investigate the relationship between diverse behavior patterns and learning performance. This study first applied the integrated method of a lag sequential analysis and sequential pattern mining together with statistical analyses (i.e., ANCOVA, nonparametric U-test) to identify sequential patterns of cognitive processes between at-risk and non-at-risk students. The results showed that the at-risk group had remained stagnant on the later days, while the non-at-risk group had developed their own strategies for how and when to effectively use the tools toward the end of gameplay. The integrated

method helped to reveal in-depth students' latent navigation behaviors and confirmed the support needs of at-risk students to develop contextual and procedural knowledge for problem-solving in the game environment. The findings of sequential analyses also inform future researchers practical guidelines when determining a method of sequential analysis.

The use of traditional educational assessments is a great challenge in understanding how students learn complex skills in solving a complex problem within open-ended serious games. Game metrics that can measure students' specific behaviors need to be developed to better understand how diverse students conduct scientific inquiry and facilitate their inquiry process in a complex learning environment such as *Alien Rescue*. This study developed game metrics (i.e., amount of new information, amount of redundant information, number of errors, and number of repeated trials) derived in the cognitive tool, Probe Design Center, and confirmed four unique groups regarding students' scientific inquiry behaviors in Probe Design Center: (1) Lack of activity group, (2) Average activity group, (3) Trial-and-error group, and (4) Best performance group. The findings suggest the needs of developing additional features as an indicator of students' scientific inquiry behavior in this serious game and of considering bigger gameplay data to build a prediction model to better understand the relationship between the game metrics and students' learning performance. Furthermore, the researcher proposed interactive visualizations of students' in-game activities—as an example of teachers' dashboard—which can support teachers to provide students real-time learning feedback as scaffolding. All in all, the integrated method of serious games analytics enabled researchers to investigate in-depth cognitive processes in the serious game to identify the challenges of students placed at-risk and their support needs. Taken together,

data-driven evidences are vital to facilitate cognitive processes in open-ended serious games with a complex problem for students with diverse characteristics.

Appendix A: Matrix of Scientific Inquiry Skills in Probe Design Center

Destination	Probe	Seismograph	Magnetometer	Thermometer	Barometer	Infrared Camera	Spectrograph
Sun	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	R	R	R	R	R	R
Mercury	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	R	R	R	R	N
Venus	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	R	N	N
Earth	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	R	R	R	R	R	R
Moon	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	N	R	N	N
Mars	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	R	N	N
Phobos	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	N	R	N	N
Deimos	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	N	R	N	N
Jupiter	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	E	R	R	R	R	N
Io	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E

	Lander	R	N	R	R	R	N
Europa	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	N	N	N
Ganymede	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	R	N	N	R	N	N
Callisto	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	R	N	R	R	N	N
Saturn	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	E	N	N	R	N	N
Titan	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	R	N	N
Uranus	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	E	N	R	R	N	N
Neptune	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	E	N	R	R	N	N
Triton	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	R	N	N
Pluto	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	R	N	N
Charon	Flyby	E	N/A	E	E	N/A	E
	Orbiter	E	N/A	E	E	N/A	E
	Lander	N	N	R	N	N	N

Note. E indicates an error, N indicates new information, R indicates redundant information, and N/A indicates a non-applicable case.

Appendix B: Space Science Knowledge Test

Name _____ Teacher _____ Period _____

Click the letter of the correct answer.

1. Which of these worlds is a planet (not a moon)?

A. Io

B. Phobos

C. Uranus

D. Not sure

2. Which of these worlds is a gas giant?

A. Saturn

B. Earth

C. Pluto

D. Not sure

3. Which of the following worlds is a moon of Jupiter?

A. Europa

B. Mars

C. Neptune

D. Not sure

4. Which of these worlds is farther from the sun than Saturn?

A. Earth's moon

B. Mercury

C. Charon

D. Not sure

5. Venus

A. is a gas giant

B. has an atmosphere denser than Earth's

C. is very cold because of a greenhouse effect

D. Not sure

6. Io

A. is the closest planet to the sun

B. has active volcanoes

C. is colder than Pluto

D. Not sure

7. Which of these worlds has the lowest surface gravity?

A. Earth

B. Triton

C. Jupiter

D. Not sure

8. What is the difference between a moon and a planet?

A. moons are closer to the sun than planets

B. planets have plant life and moons do not

C. moons orbit planets but planets do not orbit moons

D. Not sure

9. Which of the following does an atmosphere do for a world?

A. causes volcanoes to erupt

B. pushes heat out into space so the world doesn't get too hot

C. protects it from meteors

D. Not sure

10. Which of the following does a magnetic field do for a world?

A. protects it from the solar wind

B. lowers its temperature

C. gives it seasons

D. Not sure

11. Craters are caused by

A. earthquakes

B. magnetic fields

C. meteor impacts

D. Not sure

12. You are standing on the surface of a world and see the sun in the sky. The rest of the sky is black and you can see stars. What do you know about that world?

A. It is a gas giant.

B. It has no atmosphere.

C. It has no magnetic field.

D. Not sure

13. Which of the following is NOT the name of a temperature scale?

A. Fahrenheit

B. Titan

C. Celsius

D. Not sure

14. Ice

A. can be made of many substances, not just water

B. covers most of the surface of Io

C. is an element

D. Not sure

15. Which of these instruments can be used to learn about temperature on a world?

A. seismograph

B. infrared camera

C. spectrograph

D. Not sure

16. Imagine that you need to determine whether or not a moon's surface has carbon. What instrument would you use?

A. wide angle camera

B. spectrograph

C. seismograph

D. Not sure

17. Scientists want to measure the pressure of Mars' atmosphere. What instrument would they use?

A. barometer

B. thermometer

C. magnetometer

D. Not sure

18. Suppose that you want to take closeup pictures of features on the surface of Callisto, but you can only afford to send an orbiter. What instrument would you include?

A. infrared camera

B. narrow angle camera

C. barometer

D. Not sure

19. You need to design a probe to go to Titan to find out if it has a magnetic field or earthquakes. Which of the following would you choose to include on your probe?

A. a battery and a solar panel

B. a barometer and a seismograph

C. a magnetometer and a seismograph

D. Not sure

20. Scientists want to gain more accurate information about the atmosphere of Venus, especially what it's made of. What type of probe would they use and what instrument would they include?

A. an orbiter with an infrared camera

B. a flyby with a seismograph

C. a lander with a barometer

D. not sure

21. At a temperature of absolute zero

A. water melts

B. atoms stop moving

C. carbon changes from a liquid to a solid

D. not sure

22. Water boils at which of the following temperatures? (Remember to think about the different temperature scales.)

A. 32 degrees C

B. 100 degrees C

C. 100 degrees F

D. Not sure

23. Which of these could be considered a "signature" for an element?

A. a seismograph

B. an infrared picture

C. a spectrum

D. not sure

24. A world will have a magnetic field if

A. it has a thick atmosphere

B. it has liquid water

C. it has a core made of liquid metal

D. not sure

Appendix C: Solution Form Rubric

Description	Points Awarded
The student does not submit any solution.	1
The student recommends an unsuitable home for the alien species.	2
The student recommends an unsuitable home, but provides almost correct reasons to a suitable home.	3
The student recommends a suitable home, but does not provide any correct reasons to substantiate their choice.	
The student recommends a suitable home and is awarded one additional point for each reason provided to substantiate their choice.	4-8

References

AAAS Project 2061. (n.d.). AAAS Project 2061 Science Assessment Website. Retrieved from http:// assessment.aaas.org. NSF ESI-0352473; G.E. DeBoer, Principal Investigator.

Abt, C. C. (1987). *Serious games*. Lanham, MD: University Press of America (Reprint).

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of the eleventh IEEE international conference on data engineering (ICDE)* (pp. 3–14). Taipei, Taiwan.

American Association for the Advancement of Science (1993). *Benchmarks for Science Literacy*. New York: Oxford University Press.

American Association for the Advancement of Science. (2001). *Atlas of science literacy* (Vol. 1). Washington, DC.

Anderson, J. R. (1980). *Cognitive psychology and its implications.* New York: W. H. Freeman and Company.

Anderson, R. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education*, *13,* 1–2.

Anderson, E., Liu, Y.-E., Apter, E., Boucher-Genesse, F., & Popović, Z. (2010). Gameplay Analysis Through State Projection. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games* (pp. 1–8). New York, NY, USA: ACM. http://doi.org/10.1145/1822348.1822349

Andrienko G., & Andrienko N. (2008). Spatio-temporal aggregation for visual analysis of movements. In *IEEE Symp. on Visual Analytics Science and Tech.* (pp. 51–58).

Barab, S. A., Cherkes-Julkowski, M., Swenson, R., Garrett, S., Shaw, R. E., & Young, M. (1999). Principles of Self-Organization: Ecologizing the Learner-Facilitator System, *The Journal of the Learning Sciences, 8*(3&4), 349–390.

Barab, S. A., Sadler, T., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: A framework for socioscientific inquiry. *Journal of Science Education and Technology, 16*(1), 59–82.

Barrow, L. H. (2006). A Brief History of Inquiry: From Dewey to Standards. *Journal of Science Teacher Education*, *17*(3), 265–278. doi: 10.1007/s10972-006-9008-5

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013*, 11. doi:10.1155/ 2013/136864

Bera, S., & Liu, M. (2006). Cognitive tools, individual differences, and group processing as mediating factors in a hypermedia environment. *Computers in Human Behavior, 22*(2), 295-319. doi:10.1016/j.chb.2004.05.001

Bogard, T., Liu, M., & Chiang, Y. H (2013). Thresholds of Knowledge Development in Complex Problem Solving: A Multiple-Case Study of Advanced Learners' Cognitive Processes. *Educational Technology Research and Development. 61*(3), 465-503. doi:10.1007/s11423-013-9295-4

Bos, B. (2007). The effect of Texas Instrument InterActive instructional environment on the mathematical achievement of eleventh grade low achieving students. *Journal of Educational Computing Research*, *37*(4), 350-368.

Bransford, J. D., & Stein, B. S. (1993). *The Ideal Problem Solver* (2nd ed.). New York: Freeman.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review, 6,* 345-375.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141–178. doi:10.1207/s15327809jls0202_2.

Canossa, A., & Drachen, A. (2009). Patterns of play: Play-personas in user-centred game development. In *Proceedings of Breaking New Ground: Innovation in Games, Play, Practice and Theory Conference.* London: DiGRA.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*(6), 1–36.

Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development, 70*, 1098–1120.

Chi, M. T. H., & Glaser, R. (1985). Problem solving ability. In R. Sternberg (Ed.), *Human abilities: An information processing approach* (pp. 227–250). San Francisco: Freeman.

Chi, M. T. H., & Bassock, M. (1991). Learning from examples vs. self-explanations. In L. B. Resnick (Ed.), Knowing, learning, and instruction: *Essays in honor of Robert Glaser* (pp. 251–282). Hillsdale, NJ: Lawrence Erlbaum

Cho, K. L., & Jonassen, D. H. (2002). The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology: Research & Development, 50*(3), 5–2.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning and Assessment, 2*(2).

Clarke-Midura, J., Dede, C., & Norton, J. (2011). *The road ahead for state assessments*. Cambridge, MA: Policy Analysis for California Education and Rennie Center for Educational Research & Policy.

Clark, D.B., Martinez-Garza, M. M., Biswas, G., Luecht, R. M., &, Sengupta, P. (2012). Driving assessment of students' explanations in game dialog using computer-adaptive testing and hidden Markov Modeling. In Ifenthaler, D. , Eseryel, D., & Xun, G. (Eds.), *Game-based Learning: Foundations, Innovations, and Perspectives* (pp. 173–199). New York: Springer.

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. doi:10.3102/0013189X032001009.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York, NY: Chapman and Hall.

Crawford, B. A. (2014). From inquiry to scientific practices in the science classroom. In N. G. Lederman, & S. K. Abell (Eds.), *Handbook of research in science education* (pp. 515–544). NY: Routledge.

Cronbach, L. J., Linn, R. L., Brennan, R. L, & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement,* 57, 373–399.

Darling-Hammond, L., Zielezinski, M. B., & Goldman, S. (2014). *Using technology to support at-risk students' learning.* Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from https://edpolicy.stanford.edu/publications/pubs/1241

DeBoer, G. (1991). *A history of ideas in science education: Implications for practice*. New York: Teachers College Press.

DeBoer, G., Abell, C., Gogos, A., Michiels, A., Regan, T., & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. Project 2061. American Association for the Advancement of Science. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice* (pp. 231–252). Arlington, VA: NSTA Press.

de Kleer, J., & Brown, J. S. (1981). Towards a theory of qualitative reasoning about mechanisms. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dewey, J. (1910). *How we think*. Boston: D. C. Heath.

Dewey, J. (1938/1997). *Experience and education*. Macmillan.

Dewey, J. (1944). *Democracy and education: An introduction to the philosophy of education.* London: Collier-Macmillan.

Dixit, P. N., & Youngblood, G. M. (2008). Understanding playtest data through visual data mining in interactive 3D environments. *Proceedings of the 12th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia and Serious Games, Louisville (CGAMES 2008)*. Wolverhampton, England: University of Wolverhampton.

Djaouti, D., Alvarez, J., Jessel, J.-P., & Rampnoux, O. (2011). Origins of serious games. In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious games and edutainment applications* (pp. 25–43). London: Springer. doi:10.1007/978-1-4471-2161-9_3

Drachen, A., & Canossa, A. (2009). Towards gameplay analysis via gameplay metrics. In Proceedings from the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era (pp. 202–209). ACM. doi:10.1145/1621841.1621878

Drachen, A., Thurau, C., Togelius, J., Yannakakis, G. N., & Bauckhage, C. (2013). Game data min- ing. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 205–253). London: Springer.

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology and Society, 24*(3), 177–181. doi: 10.1177/ 0270467604264992

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision: Expertise in real world contexts. *Organization Studies, 26*(5), 779–792. doi: 10.1177/ 0170840605053102

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58*, 1–113.

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science, 5*, 94 –98.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K08*. Washington, DC: The National Academies Press.

English, L. D. (1998). Children's problem posing within formal and informal contexts. *Journal for Research in Mathematics Education, 29*(1), 83–106.

Fan, X., Miller, B. C., Park, K.-E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An Exploratory Study about Inaccuracy and Invalidity in Adolescent Self-Report Surveys. *Field Methods*, *18*(3), 223–244. doi: 10.1177/152822X06289161

Foundation of American Scientists. (2006). *Summit on educational games: Harnessing the power of video games for learning*. Washington, DC.

Frick, T., Myers, R., Thompson, K. & York, S. (2008). *New ways to measure systemic change: Map & Analyze Patterns & Structures Across Time (MAPSAT)*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Orlando, FL.

Frick, T., Myers, R., Howard, C. & Barrett (2011). *Applications of MAPSAT in educational research: Map & Analyze Patterns & Structures Across Time*. Paper presented at the annual conference of the Association of Educational Communications and Technology, Jacksonville, FL.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy* (2nd ed.). New York: Palgrave/Macmillan.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of Problem-Based Learning: A Meta-Analysis From the Angle of Assessment. *Review of Educational Research, 75*(1), 27–61. https://doi.org/10.3102/00346543075001027

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist, 21*, 99–120.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction, 1*(2), 129–144.

Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C. D., & Roberts, J. C. (2011). Visual comparison for information visualization. *Information Visualization, 10*(4), 289–309. doi:10.1177/1473871611416549

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining. *Journal of the Learning Sciences*, *22*(4), 521–563. doi: 10.1080/10508406.2013.837391

Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity*.

Gott, R., Duggan, S., & Roberts, R. (2008). *Concepts of evidence*. School of Education, University of Durham, UK.

Gott, R., & Murphy, P. (1987). *Assessing investigation at ages 13 and 15: Assessment of Performance Unit Science Report for Teachers: 9*. London: Department of Education and Science.

Hämäläinen, W. and Vinni, M. (2010). Classifiers for educational technology. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (Eds.), *Handbook of Educational Data Mining*, (pp. 54–74). CRC Press.

Hammons-Bryner, S. (1995). Interpersonal relationships and African American womens educational achievement. *An Ethnographic Study, 9*(1), 10-17.

Harms, N., & Yager, R. (1981). *What research says to the science teacher* (Vol. 3). Washington, DC: National Science Teachers Association.

Harpstead, E., MacLellan, C. J., Aleven, V., & Myers, B. A. (2015). Replay Analysis in Open-Ended Educational Games. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement* (pp. 381–399). Switzerland: Springer. doi: 10.1007/978-3-319-05834-4

Hélie, S. & Sun, R. (2010). Incubation, insight, and creative problem solving: A unifiedtheory and a connectionist model. *Psychological Review, 117*, 994–1024.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235-266.

Hmelo-Silver, C. E., & Azevedo, R. (2006). Understanding complex systems: Some core challenges. *Journal of the Learning Sciences*, *15*, 53–61.

Hou, H. T. (2012). Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG). *Computers & Education*, *58*(4), 1225–1233.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264-323. doi: 10.1145/331499.331504

Jonassen, D. H. (2000). Toward a design theory of problem-solving. *Educational Technology Research and Development, 48*(4), 63–85. doi: 10.1007/bf02300500

Jonassen, D. H. (2004). *Learning to Solve Problems : An Instructional Design Guide*. San Francisco: Pfeiffer.

Jonassen, D. H., & Kwon, H. I. (2001). Communication patterns in computer-mediated vs. face-to-face group problem solving. *Educational Technology Research and Development, 49*(10), 35–52.

Kamarainen, A. M., Metcalf, S., Grotzer, T., Browne, A., Mazzuca, D., Tutwiler, M. S., & Dede, C. (2013). EcoMOBILE: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education, 68*, 545–556. doi:10.1016/j.compedu.2013.02.018

Kang, J., & Liu, M. (2016). *Examining students' learning behaviors during the problem-solving process in a serious game: A prediction study*. Paper presented at annual meeting of 2016 American Educational Research Association (AERA), Washington, DC.

Kang, J., & Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior, 72*, 757-770. doi: 10.1016/j.chb.2016.09.062

Kang, J., Liu, S., & Liu, M. (2017). Tracking students' activities in serious games. In F. Lai (Eds.), *Proceedings of AECT-LKAOE 2015 Summer International Research Symposium,* Shanghai, China.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* Wiley, New York

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Eductional Data Mining, 4,* 144-182.

Kiili, K. (2007). Foundation for problem-based gaming. *British Journal of Educational Technology, 38*(3), 394-404.

Kim, J. M., & Lee, W. G. (2011). Assistance and possibilities: Analysis of learning-related factors affecting the online learning satisfaction of underprivileged students. *Computers & Education, 57,* 2395–2405.

Kim, C., Lim, J.-H., Ng, R. T., & Shim, K. (2007). SQUIRE: Sequential pattern mining with quantities. *Journal of Systems and Software, 80*(10), 1726–1745. http://doi.org/10.1016/j.jss.2006.12.562

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–55.

Kohler, W. (1925). *The mentality of apes.* New York, NY: Liveright.

Kuosa, K., Distante, D., Tervakari, A., Cerulo, L., Fernández, A., Koro, J., & Kailanto, M. (2016). Interactive Visualization Tools to Improve Learning and Teaching in Online Learning Environments. *International Journal Of Distance Education Technologies*, *14*(1), 1-21. doi:10.4018/IJDET.2016010101

Lajoie, S. P. (1993). Computer environments as cognitive tools for enhancing learning. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 261–288). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lederman, J. S., Lederman, N. G., Bartos, S. A., Bartels, S. L., Meyer, A. A., & Schwartz, R. S. (2014). Meaningful assessment of learners' understandings about scientific inquiry-The views about scientific inquiry (VASI) questionnaire. *Journal of Research in Science Teaching, 51*(1), 65–83.

Linek, S. B., Öttl, G., & Albert, D. (2010). Non-invasive data tracking in educational games: Combination of logfiles and natural language processing. In L. G. Chova, D. M. Belenguer (Eds.), *INTED 2010: International Technology, Education and Development Conference*, Spain, Valenica.

Liu, M., & Bera, S. (2005). An analysis of cognitive tool use patterns in a hypermedia learning environment. *Educational Technology Research and Development, 53*(1), 5–21. doi:10.1007/BF02504854

Liu, M., Horton, L. R., Corliss, S. B., Svinicki, M. D., Bogard, T., Kim, J., et al. (2009). Students' problem solving as mediated by their cognitive tool use: A study of tool use patterns. *Journal of Educational Computing Research, 40*(1), 111–139.

Liu, M., Kang, J., Lee, J, Winzeler, E., & Liu, S. (2015). Examining through visualization what tools learners access as they play a serious game for middle school science. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.) *Serious Games Analytics*: *Methodologies for Performance Measurement, Assessment, and Improvement* (pp. 181-208). Switzerland: Springer. doi: 10.1007/978-3-319-05834-4

Liu, M., Lee, J., Kang, & Liu, S. (2016). What we can learn from the data: a multiple-case study examining behavior patterns by students with different characteristics in using a serious game. *The Technology, Knowledge and Learning, 21*(1), 33-57. 10.1007/s10758-015-9263-7

Livingston, E., & Borko, H. (1989). Expert-novice differences in teaching: A cognitive analysis and implications for teacher education. *Journal of Teacher Education, 40*(4), 36–42. doi:10.1177/002248 718904000407.

Loh, C. S. (2006). Designing online games assessment as ‘‘Information Trails’’. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Games and simulation in online learning: Research and development frameworks* (pp. 323–348). Hershey, PA: Idea Group Inc.

Loh, C. S. (2012). Information trails: In-process assessment of game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 123–144). New York: Springer. doi:10.1007/ 978-1-4614-3546-4

Loh, C. S., Anantachai, A., Byun, J. H., & Lenox, J. (2007). Assessing what players learned in serious games: *In situ* data collection, information rails, and quantitative analysis. In Q. Mehdi (Ed.), *Computer Games: AI, Animation, Mobile, Educational & Serious Games Conference, Louiseville* (CGAMES 2007) (pp. 10–19). Wolverhampton, England: University of Wolverhampton.

Loh, C. S., & Sheng, Y. (2014). Maximum similarity index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior, 39*, 322–330.

Loh, C. S., & Sheng, Y. (2015a). Measuring Expert Performance for Serious Games Analytics: From Data to Insights. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement* (pp. 101–134). Switzerland: Springer. doi: 10.1007/978-3-319-05834-4

Loh, C. S., & Sheng, Y. (2015b). Measuring the (dis-)similarity between expert and novice behaviors as serious game analytics. *Education and Information Technologies, 20*(1), 5-19. doi:10.1007/s10639-013-9236-y

Loh, C. S., Sheng, Y,, & Ifenthaler, D. (2015a). Serious games analytics: Theoretical framework. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 3–29). Switzerland: Springer. doi:10.1007/978-3-319-05834-4.

Loh, C. S., Sheng, Y,, & Li, I. (2015b). Predicting expert-novice performance as serious game analytics with objective-oriented and navigational action seequences. *Computers in Human Behavior, 49*, 147–155. doi: 10.10116/j.chb.2015.02.053

Mayer, R. E. (2008). *Learning and instruction*. Upper Saddle River, NJ: Merrill Prentice Hall.

Mayer, R. E. (2013). Problem-solving. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 769–779). Oxford University Press.

McFarlane, A., Sparrowhawk, A., & Heald, Y. (2002). *Report on the educational use of games*. Cambridge, UK: TEEM Ltd.

Michael, D., & Chen, S. (2006). *Serious games: Games that educate, train, and inform*. Boston: Thomson Course Technology PTR.

Myers, R. D., & Frick, T. W. (2015). Using Pattern Matching to Assess Gameplay. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics* (pp. 435–458). Springer International Publishing.

National Research Council (1996). *National cience Education Standard*. Washington, DC: The National Academies Press.

National Research Council. (1997). *Introducing the National Science Standards.* Washington, DC: National Academy Press.

National Research Council (2000). *Inquiry and the National Science Education Standard*. Washington, DC: The National Academies Press.

National Research Council. (2010). *Exploring the Intersection of Science Education and 21st Century Skills: A Workshop Summary.* M. Hilton, Rapporteur. Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nelson, B., Ketelhut, D. J., Clarke, J., Dieterle, E., Dede, C., & Erlandson, B. (2007). Robust design strategies for scaling educational innovations: The River City MUVE case study. In B. E. Shelton & D. A. Wiley (Eds.), *The educational design and use of computer simulation games.* Rotterdam, The Netherlands: Sense Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Emglewood Cliffs, NJ: Prentice Hall.

Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation, 8*(2).

Partnership for 21st Century Skills (2007). *P21 Framework definitions*. Retrieved from http://www.p21.org/our-work/p21-framework

Perkins, D. (1986). Knowledge as design. Hillsdale, NJ: Erlbaum.

Physical Science Study Committee. (1960). *Physics*. Boston: Heath.

Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.

Pohl, M., Wallner, G., & Kriglstein, S. (2016). Using lag-sequential analysis for understanding interaction sequences in visualizations. International Journal of Human-Computer Studies, 96, 54-66. doi: 10.1016/j.ijhcs.2016.07.006

Ponticell, J. (2001). Making school more rewarding: At-risk students' perspectives on teaching and learning. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Quellmalz, E., Timms, M., & Schneider, S. (2009). Assessment of student learning in science simulations and games. *Proceedings of the Workshop on learning science: Computer games, simulations, and education*. Washington, DC: National Academy of Sciences.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics: Actionable measures of multidimensional learning trajectories. *British Journal of Educational Technology, 46*(1), 98–122.

Resnick, L.B., (1985). Cognition and instruction: Recent theories of human competence. In B.L. Hammonds (Ed.), *Psychology and learning: The master lecture series* (Vol. 4, pp. 127–186). Washington D.C.: American Psychological Association.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27. doi:10.1002/widm.1075.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. D. (Eds.). (2010). *The handbook of educational data mining* (1st ed.). Boca Raton, FL: CRC Press.

Samsonov P., Pedersen S., Hill C. L. (2006). Using problem-based learning software with at-risk students: A case study. *Computers in the Schools, 23*(1), 111–124.

Sawyer, B. (2009). Foreword: From virtual U to serious game to something bigger. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. xi–xvi). New York, NY: Routledge.

Sawyer, B., & Rejeski, D. (2002). *Serious games: Improving public policy through game-based learning and simulation*. Washington, DC.

Scarlatos, L. L., & Scarlatos, T. (2010) *Visualizations for the assessment of learning in computer games.* Paper presented at the 7th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT 2010), Incheon, Korea.

Schmidt, R. A., & Lee, T. (2011). *Motor control and learning: A behavioral emphasis* (5th ed.). Champaign, IL: Human Kinetics.

Schunk, D. H. (2016). *Learning theories: an educational perspective*. Upper Saddle River, NJ: Pearson.

Seif El-Nasr, M., Drachen, A., & Canossa, A. (Eds.). (2013). *Game Analytics*. London: Springer London. Retrieved from http://link.springer.com/10.1007/978-1-4471-4769-5

Schell, J. (2008). *The art of game design: A book of lenses* (1st ed.). Burlington, MA: Morgan Kaufmann.

Schraw, G., Dunkle, M.E., & Bendixen, L.D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology, 9,* 523–538.

Schwab, J. J. (1966). *The teaching of science*. Cambridge, MA: Harvard University Press.

Sanderson, P.M., Fisher, C. (1994). Exploratory sequential data analysis: foundations. *Human Computer Interactation, 9*(4), 251–317.

Siemens, G. (2013). Learning Analytics The Emergence of a Discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. doi: 10.1177/0002764213498851

Simons, K., & Klein, J. (2007, January). The impact of scaffolding and student achievement levels in a problem-based learning environment. *Instructional Science, 35*(1), 41–72. doi:10.1007/s11251-006-9002-5.

Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A Meta-Analysis of Data Collection in Serious Games Research. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement* (pp. 31–55). Switzerland: Springer. doi: 10.1007/978-3-319-05834-4

Spires, H., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem solving and game-based learning: effects of middle grade students' hypothesis testing strategies on science learning outcomes. *Journal of Educational Computing Research, 44*, 453–472.

Spring, F., & Pellegrino, J. W. (2011). The challenge of assessing learning in open games: HORTUS as a case study. *Proceedings of the 8th Games+Learning+Society Conference—GLS 8.0* (pp. 209–217).

Squire, K. (2008). Open-ended video games: A model for developing learning for the interactive age. In K. Salen (Ed.) *The ecology of games: Connecting youth, games, and learning.* (167–198). Cambridge, MA: The MIT Press. doi:10.1162/dmal. 9780262693646. 167.

Stecher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1–14.

Stevens J. P. (2009). *Applied Multivariate Statistics for the Social Sciences* (4th ed.). New York: Routledge

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institute.

Texas Education Agency. (2017).

Thorndike, E. L. (1913). *Educational psychology: Vol. 1. The original nature of man*. New York: Teacher's College Press.

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). *Analytics in higher education: Establishing a common language*. EDUCAUSE Learning Initiative.

van Merriënboer, J. J. (2013). Perspectives on problem-solving and instruction. *Computers & Education, 64*, 153–160. doi: 10.1016/j.compedu.2012.11.025

Wallner, G., & Kriglstein, S. (2013). Visualization-based analysis of gameplay data—A review of literature. *Entertainment Computing, 4*(3), 143–155. doi:10.1016/j.entcom.2013.02.002.

Wallner, G., & Kriglstein, S. (2015). Comparative visualization of player behavior for serious game analytics. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.) *Serious Games Analytics*: *Methodologies for Performance Measurement, Assessment, and Improvement* (pp. 159–179). Switzerland: Springer. doi: 10.1007/978-3-319-05834-4

Wallas, G. (1926). *The Art of Thought.* New York: Harcourt Brace.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development, 53*(4), 5-23.

Wecker, C., Rachel, A., Heran-Dorr, E., Waltner, C., Wiesner, H., & Fischer, F. (2013). Presenting theoretical ideas prior to inquiry activities fosters theory-level knowledge. *Journal of Research in Science Teaching, 50*(10), 1180–1206.

Welch, W. W., Klopfer, L. E., Aikenhead, G. S., & Robinson, J. (1981). The role of inquiry in science education: Analysis and recommendations. *Science Education, 65*(1), 33–50.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory and Cognition, 26*(4), 716–730.

Zhou, M., Xu, Y., Nesbit, J. C., & Winne, P. H. (2010). Sequential pattern analysis of learning logs: Methodology and applications. In C. Romero et al. (Eds.), *Handbook of Educational Data Mining* (pp. 107–121). Chapman & Hall/CRC Press.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M.Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic Press.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25–32. doi:10.1109/MC.2005.297